

MICHIGAN STATE TESTING PROGRAM

MICHIGAN ENGLISH LANGUAGE PROFICIENCY ASSESSMENT (MI-ELPA)

TECHNICAL MANUAL: 2006 ADMINISTRATION

KINDERGARTEN THROUGH GRADE 12

SUBMITTED TO THE
MICHIGAN STATE DEPARTMENT OF EDUCATION

FEBRUARY 2007

TABLE OF CONTENT

TABLE OF CONTENT	i
FIGURES AND TABLES	iii
OVERVIEW OF THIS MANUAL.....	1
SECTION 1. INTRODUCTION.....	3
1.1 Background	3
1.2 Rationale and Purpose	3
1.3 Recommended Test Use.....	4
1.4 Test Accommodations.....	4
Large Type	5
Braille.....	5
SECTION 2. TEST DESIGN AND DEVELOPMENT	6
2.1 Overview.....	6
2.2 Test Specifications by Modality and Grade Span.....	6
2.3 Item Blueprints by Michigan Learning Standards by Grade Span, Modality, and Form.....	7
2.4 Item Development and Review Processes.....	7
Differential Item Functioning (DIF).....	8
2.5 Test Construction.....	10
Testing Written Language	11
Testing Oral Language.....	11
SECTION 3. SCORING	13
3.1 MI-ELPA Range Finding.....	13
3.2 Rater Training	14
3.3 Inter-Rater and Intra-Rater Reliability	14
3.4 Calibration Sets	15
3.5 Monitoring Reports	15
3.6 Retraining.....	16
SECTION 4. CLASSICAL ITEM-LEVEL AND MODALITY (SUBTEST) STATISTICS.....	17
4.1 Classical Test Theory	17
4.2 Item-Level Descriptive Statistics	18
4.3 Measure of Central Tendency.....	18
4.4 MI-ELPA Stand-Alone Field-Test Item Statistics	21
SECTION 5. RELIABILITY.....	23
5.1 Internal Consistency Reliability	23
5.2 Classical SEM (based on Classical Test Theory).....	23
5.3 Conditional SEM (based on Item Response Theory)	24
5.4 Inter-Rater Reliability	25
5.5 Reliability of Each of the Five Modalities	25
5.6 Reliability of Classification Decision at Proficient Cut	28
SECTION 6. VALIDITY	31
6.1 Test Content.....	31
6.2 Evidence of the Test Content for the MI-ELPA	31
6.3 Internal Structure	32
6.4 Evidence of the Internal Structure of the MI-ELPA.....	32
6.5 Relationships to Other Variables	35
Performance Differences Between Native and Non-Native English-Speaking Students Taking the SELP ...	35

2006 MI–ELPA Technical Manual

<i>Relationship Between the SELP and the Stanford Diagnostic Reading Test (SDRT)</i>	36
<i>Relationship Between the SELP and the Abbreviated Reading Subtest of the Stanford Achievement Test Series, Ninth Edition (Stanford 9)</i>	36
SECTION 7. CALIBRATION, EQUATING, AND SCALING	37
7.1 <i>The Rasch and Partial Credit Models</i>	37
7.2 <i>Calibration, Equating, and Scaling of the MI-ELPA</i>	39
7.3 <i>Vertical Scaling of SELP</i>	40
<i>Forms Equating</i>	41
<i>Scale Scores</i>	42
7.4 <i>Linking MI-ELPA Scale to the SELP Vertical Scale</i>	42
7.5 <i>Scale Scores for the MI-ELPA</i>	43
7.6 <i>Test Characteristic Curves for the MI-ELPA by Grade Span</i>	44
7.7 <i>Linking Subsequent MI-ELPA Operational Tests Across Years</i>	45
SECTION 8. IRT STATISTICS	46
8.1 <i>Model and Rationale for Use</i>	46
8.2 <i>Evidence of Model Fit</i>	46
8.3 <i>Rasch Statistics</i>	48
8.4 <i>Item Information</i>	49
SECTION 9. STANDARD SETTING	50
9.1 <i>Introduction</i>	50
9.2 <i>Standard-Setting Methods</i>	50
9.3 <i>Standard-Setting Model and Process</i>	51
9.4 <i>Committees of Panelists</i>	52
9.5 <i>Performance Levels and Cut-Scores</i>	53
9.6 <i>The Use of the Vertical Scale</i>	54
9.7 <i>Standard-Setting Process</i>	55
<i>Review of the Assessment</i>	55
<i>Experiencing the Assessment</i>	55
<i>Scoring the Assessment</i>	56
<i>Review of Student Performance Levels</i>	56
<i>Three Rounds of Ratings</i>	56
<i>Evaluation</i>	57
9.8 <i>Agendas</i>	57
9.9 <i>Summary Statistics for the Three Rounds of Ratings</i>	57
9.10 <i>Evaluation Results</i>	59
9.11 <i>Post-Standard-Setting Analyses</i>	60
9.12 <i>Final Performance-Level Cut-Scores for the MI-ELPA</i>	60
9.13 <i>Calculation of Achievement "Targets" for Each Modality</i>	61
9.14 <i>Calculation of the Performance-Level Cuts for the Screener</i>	61
SECTION 10. SUMMARY OF OPERATIONAL TEST RESULTS	62

FIGURES AND TABLES

<i>Table 2.1: Test Specifications by Modality and Grade Span.....</i>	<i>7</i>
<i>Table 2.2: Maximum Number of Points by Modality and Grade Span.....</i>	<i>7</i>
<i>Table 2.3: Classification of DIF for Dichotomously and Polytomously Scored Items</i>	<i>10</i>
<i>Table 4.1: Summary Statistics of MI-ELPA Modalities by Grade Span</i>	<i>19</i>
<i>Table 4.2: Summary Statistics of MI-ELPA Modalities by Grade.....</i>	<i>20</i>
<i>Table 5.1: Descriptive Statistics and Reliability by Grade and Modality</i>	<i>26</i>
<i>Figure 5.1: Classification Accuracy.....</i>	<i>28</i>
<i>Figure 5.2: Classification Consistency.....</i>	<i>28</i>
<i>Table 5.2: Decision and Consistency Table by Grade.....</i>	<i>30</i>
<i>Table 6.1: Intercorrelations Among Modalities by Grade</i>	<i>33</i>
<i>Figure 7.1: Category Response Curves for a Two-Step Item Using the PCM.....</i>	<i>38</i>
<i>Table 7.1: Equating of Levels Research Design.....</i>	<i>41</i>
<i>Table 7.2: Scale Score Transformation Equations for MI-ELPA for Total Test and Modalities</i>	<i>43</i>
<i>Figure 7.2: MI-ELPA Test Characteristic Curves (TCCs) by Grade Span</i>	<i>44</i>
<i>Table 8.1: Criteria to Evaluate Mean-Square Fit Statistics</i>	<i>47</i>
<i>Table 8.2: Average Rasch Difficulty by Grade Span and Modality.....</i>	<i>48</i>
<i>Table 9.1: Panel Composition for Standard-Setting Committees.....</i>	<i>53</i>
<i>Table 9.2: Primary School Level Raw Score Standards by Rounds</i>	<i>58</i>
<i>Table 9.3: Elementary School Level Raw Score Standards by Rounds</i>	<i>58</i>
<i>Table 9.4: Middle School Level Raw Score Standards by Rounds</i>	<i>58</i>
<i>Table 9.5: High School Level Raw Score Standards by Rounds.....</i>	<i>59</i>
<i>Table 9.6: Final Performance-Level Cut-Scores.....</i>	<i>60</i>
<i>Table 10.1: Raw-Score Summary by Grade, Modality, and Total Test</i>	<i>62</i>
<i>Table 10.2: Scale-Score Summary.....</i>	<i>64</i>
<i>Table 10.3: Percent of Students in Each Proficiency Level By Grade</i>	<i>64</i>
<i>Table 10.4: Percent of Students in Each Proficiency Level by Grade and Modality.....</i>	<i>65</i>

OVERVIEW OF THIS MANUAL

This Michigan English Language Proficiency Assessment (MI-ELPA) Technical Manual for the 2006 administration is organized around ten major sections—Introduction; Test Design and Development; Scoring; Classical Item-Level and Subtest (Modality) Statistics; Reliability; Validity; Calibration, Equating, and Scaling (CES); Item Response Theory (IRT) Statistics; Standard Setting; and Summary of Operational Test Results. An overview of this manual is provided below.

Section 1

This section presents the background, rationale, purpose, recommended test use, and test accommodations. Test accommodations include large type and Braille.

Section 2

This section describes the test development process of the MI-ELPA. It includes the test specifications and the item development and review processes, including differential item functioning (DIF) analysis, item field testing of the Writing subtest, and test construction.

Section 3

This section provides a description of the scoring process. It includes the description of the range-finding meeting that was held in Lansing, Michigan. It also provides information about results of the inter-rater and intra-rater reliability and the rater agreement analyses.

Section 4

This section begins with a brief description of the Classical Test Theory (CTT), followed by item-level descriptive statistics based on CTT.

Section 5

This section explains internal consistency reliability, classical Standard Error of Measurement (SEM), and conditional SEM based on IRT. It also provides the reliability of each of the four modalities and the reliability of classification decision at the proficient cut.

Section 6

This section describes the validity studies that were conducted. It includes evidence of validity based on test content, internal structure, and relationships to other variables.

Section 7

This section explains the Rasch and Partial Credit Models and provides sample item characteristic curves for a one-step item and a two-step item. It also includes the results of the calibration, equating, and scaling of the 2006 administration of the MI-ELPA.

Section 8

This section explains the rationale for use of the IRT model. It includes the IRT model fit statistics and the average Rasch difficulty of the subtests.

Section 9

This section presents the standard-setting process that was followed to establish the performance-level cuts. It includes the standard-setting model, the standard-setting process, summary statistics for the round-by-round ratings, evaluation results, post-standard-setting analyses, and final performance-level cut-scores.

Section 10

This section presents the raw score summary, scale score summary, and percentage of students in each performance category for the 2006 administration of the MI-ELPA.

SECTION 1. INTRODUCTION

1.1 Background

Title III of the federal No Child Left Behind (NCLB) Act of 2001 requires annual assessment of the English proficiency of limited English proficient students. NCLB requires demonstrated annual improvement and adequate yearly progress for such students in order for them to develop English proficiency and meet challenging state academic content and student achievement standards. The state of Michigan’s Office of Educational Assessment and Accountability (OEAA) regulations also require annual assessment of limited English proficient students using a state-approved assessment.

To meet these requirements, OEAA requested test development, research, and scoring for the four grade spans and four language modalities that form the framework of Michigan’s approved English as a Second Language (ESL) learning standards. The test was developed for grade spans K–2, 3–5, 6–8, and 9–12, with Speaking, Listening, Reading, and Writing modalities at each grade span, to assess the English language proficiency of students in kindergarten through grade 12 who are English language learners. Proficiency on a fifth modality, Comprehension, which was a composite of selected items from Listening and Reading, was also assessed. The test was developed in accordance with the *Standards for Educational and Psychological Testing* (American Educational Research Association, 1999) and OEAA testing requirements. The test is consistent with the principles of universal design and also consistent with applicable federal and state testing requirements.

In response to the challenging timeline that OEAA presented, Harcourt and OEAA agreed to a creative and robust solution with two distinct phases. For the first phase, Harcourt used content from the Harcourt English language learner item bank, as well as Mountain West Assessment Consortium (MWAC) and several Michigan Educational Assessment Program (MEAP) items, to produce custom forms for the 2006 test administration. The 2006 administration also included new embedded field-test items. For the second phase, more new field-test items and fewer Harcourt item bank and Mountain West items will be used in order to produce custom forms beginning with the 2007 test administration.

1.2 Rationale and Purpose

OEAA has established learning standards for all English language learners (ELLs) attending Michigan schools. In compliance with NCLB, which mandates that all ELLs from kindergarten through grade 12 be assessed every year to measure their English language proficiency in listening, speaking, reading, and writing and that their annual progress toward proficiency be tracked, OEAA developed an annual test that measures student progress toward meeting the state’s standards. This test is the MI-ELPA. The MI-ELPA helps schools determine which instructional standards teachers must focus on to ensure their ELLs fully acquire the language proficiency that will prepare them for success in the classroom.

The purpose of the test is to measure annual student improvement in achieving English language proficiency in order to ultimately exit an ESL or bilingual education program, move into an English Language Arts classroom, and function successfully without any additional support. The test also provides “targets” for students to achieve in each of the four modalities of the MI–ELPA, Speaking, Listening, Reading, and Writing. The targets are simply the proficiency-level expectations of students in a particular modality. A fifth modality, Comprehension, is also included in the test, though the Comprehension modality does not have specific targets identified since this modality is composed of items selected from the Listening and the Reading modalities.

Furthermore, OEAA, in conjunction with Harcourt, has developed shorter versions of the total tests to determine the placement of students in the ELL program. These Screener tests comprise selected items from the total MI–ELPA and Harcourt item bank and are developed for testing students at levels K, 1–2, 3–5, 6–8, and 9–12. The performance levels for passing the Screeners are the same as the achievement ability required to pass the MI–ELPA.

1.3 Recommended Test Use

The MI–ELPA is designed to assess students at all proficiency levels within each grade span. This vertical development of the language tested allows the test to discriminate more finely among students at different stages of language acquisition. Because test results provide students, teachers, and parents with an objective report of each student’s strengths and weaknesses in the English language skills of listening, speaking, reading, and writing, the MI–ELPA helps determine whether these students are making adequate progress toward English language proficiency. Year-to-year progress in language proficiency can also be measured and documented after the MI–ELPA vertical scale is successfully established. The test results can also help schools focus on ways to improve instruction so that ELLs become proficient in English, thereby freeing them to focus on content-based materials, such as mathematics and science.

As explained in the above section, the Screeners are designed to assess students’ need for enrollment in the ELL program.

1.4 Test Accommodations

All test items were developed following the guidelines of universal design. Adherence to these guidelines ensured that the assessments were accessible and valid for the widest range of students, including students with disabilities. Applying universal design during the development process helped eliminate the need to address after-the-fact accommodations and provided a better assessment for all students. Checklists were used to review every item to ensure it was built while taking into consideration equitable use, flexibility in use, simple unintuitive design, perceptible information tolerance for error, low physical effort and size and span for approach and use. During forms construction, Harcourt utilized in-house content and fairness experts to ensure that the forms were pulled with concepts of universal design in mind. Harcourt stringently reviewed forms for special populations—such as visually- or hearing-impaired students—to ensure that items were fair, reliable, and accessible to all.

Large Type

Harcourt has standardized large-type product specifications that serve to ease the test-taking experience for students who require large type. One form in large type per grade span, with type 18 points minimum and not larger than 24 points for titles, was produced. Pages are printed in black only and on a cream colored, non-glare vellum stock to ease readability of pages. Covers are printed on heavier stock to provide stiffness to the booklets, which protects interior text pages. Plastic spiral binding makes turning of pages easy to accomplish.

Braille

Harcourt created the Braille version of the MI-ELPA using certified and experienced transcribers who can deal with the multiple codes, rules, and guidelines. Harcourt produced Braille forms for each MI-ELPA subtest and grade span. For the K–2 level, a checklist was provided rather than a Braille test.

If an area was difficult to Braille, Harcourt determined with content specialists whether there were other ways that the construct could be worded or measured. If, however, an item was impossible to adapt for students with vision problems, it was dropped from the Braille version of the MI-ELPA.

SECTION 2. TEST DESIGN AND DEVELOPMENT

2.1 Overview

To meet the requirements of Title III of the federal No Child Left Behind (NCLB) Act and of Michigan regulations regarding the assessment of limited English proficient students, OEAA requested test development, research, and scoring for the four grade spans and four language modalities that form the framework of Michigan’s approved English as a Second Language (ESL) learning standards. The test was developed for four grade spans (K–2, 3–5, 6–8, 9–12) and in four modalities (Speaking, Listening, Reading, and Writing) to assess the English language proficiency of students in kindergarten through grade 12 who are ELLs. Proficiency on a fifth modality, Comprehension, which was a composite of selected items from Listening and Reading, was also assessed. The test was developed in accordance with the *Standards for Educational and Psychological Testing* (American Educational Research Association, 1999) and the OEAA testing requirements. The test is consistent with the principles of universal design and also consistent with applicable federal and state testing requirements.

2.2 Test Specifications by Modality and Grade Span

The MI-ELPA includes a total of five modalities (Speaking, Listening, Reading, Writing, and Comprehension) for grades K–12. It includes multiple-choice, constructed-response, short-response, and extended-response items. The total number of items per grade span varies. For grade span K–2 there are a total of 69 items, for grade span 3–5 there are a total of 72 items, for grade span 6–8 there are a total of 74 items, and for grade span 9–12 there are a total of 80 items.

The Speaking modality has 13 constructed-response items for all grade spans except 3–5, where there are 12. The Listening and Reading modalities both consist of multiple-choice items only. The number of items for the Listening modality varies from 21–24 for the different grade spans. The number of items for the Reading modality varies from 21–25 for the different grade spans. The number of items for the Writing modality ranges from 13–18 for the various grade spans. The Writing modality comprises three parts:

- Multiple-choice section that assesses ELLs’ understanding of the principles of written English at the phoneme, word, and sentence levels
- Developmental writing items (K–2 only)
- Sentence writing items, paragraph writing items, and items requiring extended response to a graphics-based prompt (number and type vary by grade span)

The test design for the 2006 administration of the MI-ELPA is shown in Table 2.1. Table 2.2 provides the maximum number of points by modality by grade span. This design consists of items from the Harcourt ELL item bank, MWAC, and MEAP.

Table 2.1: Test Specifications by Modality and Grade Span

Grade Span	Listening	Speaking	Reading	Writing		Comprehension	Total Number of Items Per Grade Span (MC + CR)
	MC	CR	MC	MC	CR	MC	
K–2	21	13	22	7	6	33	69
3–5	21	12	21	13	5	36	72
6–8	21	13	23	13	4	32	74
9–12	24	13	25	13	5	33	80

Note. Comprehension comprises items selected from Listening and Reading modalities and is not included in the column titled “Total Number of Items Per Grade Span.”

Table 2.2: Maximum Number of Points by Modality and Grade Span

Grade Span	Listening	Speaking	Reading	Writing		Comprehension	Total Number of Points Per Grade Span (MC + CR)
	MC	CR	MC	MC	CR	MC	
K–2	21	23	22	7	12	33	85
3–5	21	23	21	13	10	36	88
6–8	21	25	23	13	10	32	92
9–12	24	25	25	13	12	33	99

Note. Comprehension comprises items selected from Listening and Reading modalities and is not included in the column titled “Total Number of Items Per Grade Span.”

2.3 Item Blueprints by Michigan Learning Standards by Grade Span, Modality, and Form

Appendices A.1–A.4 provide in detail the item blueprints by Michigan Learning Standards by grade span, by modality, and by form.

2.4 Item Development and Review Processes

In order to create a new and fully aligned assessment for ELLs for the 2006 administration, and also to meet the reporting requirements for NCLB in 2006, Harcourt made use of a bank of field-tested ELL items and commissioned passages and stimuli. The Harcourt ELL item bank included items developed for the Stanford English Language Proficiency (SELP) Test forms. Items in the bank were originally submitted by item writers who are also educators of ELLs. Assessment specialists at Harcourt reviewed the items created, and in accordance with the item specifications, the assessment specialists ensured the following:

- Absence of bias and sensitive topics in passages
- Item soundness
- Absence of bias in items
- Appropriateness of topic, vocabulary, and language structure for each grade span
- Match to the intended Michigan State ESL standard

Each test question was rigorously reviewed by ELL educators. Only those test questions judged to be of acceptable quality and to be fair to students who come from all over the world were approved for inclusion in the item bank. The test questions were also sampled in classrooms with ELLs to ensure that the directions are clear and easy to follow and that the tests are interesting to students and are reliable indicators of student achievement. Although the tests are challenging for students, the questions, graphics, and stories engage students and reflect the kinds of activities in which they are involved on a daily basis. This helps to ensure that the tests will measure the learning of each individual student and provide meaningful information about his or her English language proficiency.

Using content from this item bank, which aligns to Michigan State ESL standards, Harcourt met the challenging spring 2006 timeline requirements for the MI-ELPA.

New items introduced for the MI-ELPA were also checked for bias. Statistically, all field-test items were analyzed for differential item functioning (DIF). Those items that showed moderate DIF were examined for the possibility of bias while those with extreme DIF were scrutinized in-depth for bias.

Differential Item Functioning (DIF)

The SELP and MWAC items had already been tested for DIF prior to their administration. The MI-ELPA had newly constructed field-test items embedded within the operational forms as well as extra field-test items that were administered as stand-alone items during fall 2006. All the field-test items were eligible for DIF testing. However, because of the small n-counts, the DIF procedure only compared students for cases in which enough student responses were available, i.e., white students with Hispanic students, and male with female students. Also the stand-alone field-test items did not have great enough n-counts for DIF analysis across ethnicity and gender. As such, DIF was only performed on the embedded field-test items. In these comparisons, white and male students were considered reference groups with respect to the comparisons for ethnicity and gender respectively.

Since the MI-ELPA included constructed-response items that were polytomously scored, the Mantel-Haenszel odds ratio α could not be used as a DIF index for all the items in the form. Instead, a generalization of the Mantel-Haenszel (1959) procedure for ordered categories, the *Mantel Statistic* (Mantel, 1963), was used for the assessment of DIF in the mixed-format examinations. The Mantel chi-square involves comparing the mean for two groups, conditional on a matching variable. It has 1 degree of freedom under the null hypothesis of no conditional association between group membership and response. For dichotomous items the Mantel statistic reduces to the usual Mantel-Haenszel chi-square statistic (without continuity correction).

The Mantel statistic has the following mathematical formulation:

$$\text{Mantel Chi-square} = \frac{\left(\sum_K F_K - \sum_K E(F_K) \right)^2}{\sum_K \text{Var}(F_K)}, \quad (1)$$

where F_K represents the sum of scores for the Focal group at the K th level of the matching variable, E represents the expected, and Var represents the variance of F_K .

$$F_K = \sum_T y_T n_{FTK}, \quad (2)$$

where y_T represents the T scores that can be obtained on the item while n_{FTK} denotes the number of focal group members who are on the k th level of the matching variable and received an item score of y_T . The expectation of F_K under the hypothesis of no association is

$$E(F_K) = \frac{n_{F+K}}{n_{++K}} \sum_T y_T n_{+TK}. \quad (3)$$

DIF statistical procedures compute the probability that one demographic group is more likely to answer an item correctly than another group, when the groups are equally able. This information is useful in reviewing items and tests for potential bias in items.

The Mantel-Haenszel and the Mantel statistic, however, offer a significance test of the presence of DIF without an indication of the direction of DIF, i.e., whether in favor of the reference or the comparison group. The statistic has low power in detecting an association in which the pattern of association for some of the strata is in the opposite direction of the patterns displayed by other strata. On the other hand, as a significance test, its power increases with the number of responses in the two groups of comparison.

For both the dichotomous and the polytomous items, *standardized mean differences* (SMD) (Zwick, Donoghue & Grima, 1993) were used as an effect type index for DIF. The SMD, which take into account the natural ordering of the item's response levels, are based on only those ability levels in which members of the comparison groups are present. This index also helps in discerning the direction of DIF. A negative value for SMD reflects DIF in favor of the reference group and against the comparison or focal group.

Mathematically SMD is defined as follows:

$$SMD = \sum_k p_{Fk} m_{Fk} - \sum_k p_{Rk} m_{Rk}, \quad (4)$$

where p_{Fk} represents the proportion of focal group members who are at the k th level of the matching variable, m_{Fk} represents the mean item score for the focal group at the k th level, and m_{Rk} represents the analogous value for the reference group.

DIF classification is indicated by the use of the Mantel statistic for polytomously scored items and the Mantel-Haenszel (MH) chi-square for the dichotomously scored items in conjunction with SMD divided by the total group standard deviation (SD)—(see Table 2.3). As shown in the table, DIF is categorized as “no DIF” (A), “mild DIF” (B) or “extreme DIF” (C).

Table 2.3 Classification of DIF for Dichotomously and Polytomously Scored Items

Type of Item	Condition of Category Classification	DIF Category
Dichotomously scored	MH Chi-square not significant or MH chi-square significant and $ SMD/SD \leq 0.17$	A
	MH Chi-square significant and $0.17 < SMD/SD \leq 0.25$	B
	MH Chi-square significant and $ SMD/SD > 0.25$	C
Polytomously scored	Mantel Chi-square not significant, or Mantel chi-square significant and $ SMD/SD \leq 0.17$	A
	Mantel Chi-square significant and $0.17 < SMD/SD \leq 0.25$	B
	Mantel Chi-square significant and $ SMD/SD > 0.25$	C

OEAA reviewed the MI-ELPA forms prior to administration. A list of all the embedded field-test items with DIF classification based on gender and ethnicity is shown in Appendices C.7-A and C.7-B respectively. The stand-alone field-test items had very small n-counts for a meaningful comparison.

2.5 Test Construction

Items selected from the Harcourt ELL item bank for the 2006 MI-ELPA represented a complete range of difficulty at all grade levels from K–12. Items ranged from very simple ones with high p -values, primarily aimed at students with very limited ability in English, to items with low p -values, aimed at students with advanced ability in English. Therefore, the number of both multiple-choice and constructed-response items was increased at each proficiency level, meeting the requirement of the OEAA.

Testing Written Language

A fundamental consideration in constructing the MI-ELPA is the language that is being tested. While this question can generally be answered from the test developer's native speaker intuition, more rigorous methods in language choice need to be applied to provide consistency across the forms of the four grade spans and to create a vertical structure within each form. By vertical structure, we mean language that ranges from the most simple, which is first acquired by non-native speakers, to advanced language that would indicate a level of English proficiency sufficient for participation in regular academic classes.

For the MI-ELPA, a test designed to assess students at all proficiency levels within each grade span, this vertical development of the language tested allows the test to discriminate more finely among students at different stages of language acquisition. Being able to accurately identify students at different levels of language development provides better information to classroom teachers, who must find the most effective way to help their students reach proficiency. It also provides the very important evidence of students' progress toward proficiency that is required by the NCLB legislation.

To determine the appropriate language for ELL items and stimuli, Harcourt assessment specialists, editors, and item and passage writers apply the Flesch-Kincaid grade-level readability measures to all reading passages. Readability measures are primarily based on factors such as the number of words in the sentences and the number of letters or syllables per word. Additionally, ESL assessment specialists also evaluate the coherence of a passage, the number of anaphora, vocabulary difficulty, sentence and text structure, and concreteness and abstractness. It is the sum of these that determines the appropriateness of the language of a passage.

There is a gradual increase in difficulty from passage to passage at every grade span, so that each form includes beginning-level passages as well as passages that are representative of on-grade reading passages found on English Language Arts reading tests. Harcourt also uses the *Educational Developmental Laboratory (EDL) Core Vocabularies in Reading, Mathematics, Science, and Social Studies*, published by Steck-Vaughn, to help determine age- and grade-appropriate language for ELL items and stimuli for the oral language subtests. Not of trivial importance is the selection of language that is topic-appropriate. Harcourt ESL assessment specialists and editors ensure that the language in all stimuli and items, from kindergarten through grade 12, is both topic- and age-appropriate for test takers.

Testing Oral Language

Recognizing that oral language structure and vocabulary of English differ vastly from the written language, issues of oral language assessment among kindergarten through grade 12 ELLs have been the subject of special investigation at Harcourt. Harcourt's ELL professionals have conducted research on the item types that appear in the MI-ELPA Listening and Speaking subtests by presenting exemplars of these item types to ELLs during cognitive labs, carefully observing and recording student responses and eliciting their reactions. Outcomes of the cognitive labs led to important design decisions regarding:

- Item types
- Number of items
- Length of pauses between items
- Use of recorded stimuli
- Recording student spoken responses

The Listening and Speaking subtests of the MI-ELPA are based on these decisions. To ensure that the language in the Listening and Speaking stimuli and items reflect current spoken language as much as possible, Listening and Speaking scripts are submitted to a read-aloud proofing process with ELL assessment specialists and editors. Additionally, for the oral components of the MI-ELPA to be relevant, the Listening and Speaking subtests must have predictive validity for academic achievement; therefore, academic language as well as social language is an integral part of the Listening and Speaking subtests of the MI-ELPA.

SECTION 3. SCORING

This section describes the open-ended scoring process for the operational and field-test items for spring 2006. Each grade span had four teams of at least five readers each. For monitoring purposes, twenty percent of each scorer’s daily output received a check score—a second reading by a team leader—as a means of tracking inter-rater reliability. Anchors, training sets, and rubrics were used as scoring guides. If questions arose during scoring, usually the problem was discussed by the group to maintain consistency in scoring.

The details of the scoring process for the operational items are described below.

3.1 MI-ELPA Range Finding

Range finding was held in San Antonio on May 15 through May 17, 2006. The participants included:

- One full-time Harcourt Supervisor and eight temporary Harcourt Performance Assessment Scoring Center (PASC) facilitators (two for each grade span: K–2, 3–5, 6–8, 9–12)
- One state department representative and four Michigan teachers made the anchor and training set decisions for each grade span.

Teachers were informed of the selection process for “paper-pulling.” At Harcourt, a team of two developers read several hundred papers to find clear-cut, typical examples of score points to share with the teachers. This range of papers also contained exemplars that would be helpful to include in training sets to make scoring clear. All developers were well-acquainted with the prompts, rubrics, and hundreds of papers reviewed during paper-pulling.

Sample responses for each item were sorted into preliminary range sets. These sets were presented to the teachers during range finding. The sets ranged from possible low to high responses and one set included a mixed range of papers. Each set included at least 15 papers.

Teachers read and assigned scores to each paper and then, as a group, discussed the scores they gave. The group came to a consensus of how each paper should be scored. After coming to agreement about the scores, the group discussed the merits of each paper and selected which would be used as anchors and which would be used for training sets. They used the rubrics as their scoring guides.

Harcourt’s PASC facilitators documented discussions and decisions made at each grade-span session. These facilitators later became the scoring trainers. All notes taken during range finding were used by facilitators during training.

The anchor sets contained three examples of each score point. Training sets included papers that helped discriminate between “line papers.” A variety of examples was used to show other types of responses different from the anchors, as well as those similar to anchor papers.

Papers selected were carefully reviewed and compared through this process to ensure consistency.

For the 0–4 scale writing prompts, Harcourt SELP items and previously developed anchors and training sets were used for training.

3.2 Rater Training

The accuracy of scoring was monitored by room directors who served as trainers for each grade span. These trainers are seasoned PASC readers who have vast experience in all facets of scoring. They carefully monitor the scoring and accuracy of their teams of readers. The room directors for this project were also the developers of the training sets and facilitated range finding.

Prior to scoring, each room director conducts team leader training. Team leaders are the next-level experts for the items being scored as well as for the requirements and procedures for the project. Each grade span had two room directors with two teams each. Team leaders went through the same training received by the readers. They actually went through training twice; once during their own session and the next with the readers the following day. Logistics of the scoring sessions and routines were discussed. This included the standards for qualifying to score, monitoring for accuracy and reliability, and procedures for retraining and evaluating readers on their teams.

All PASC readers have a minimum of a bachelor’s degree and have successfully completed generalized workshops in performance assessment scoring before ever being considered as a potential for a specific project, such as the MI-ELPA. Training of readers was conducted by the room director for each grade span and is based on the anchors and training sets developed by Michigan teachers during range finding. After training, each reader was given qualifying sets (mini-tests) to apply the criteria they had learned. Two attempts were possible. Any reader who failed to meet the standard at this time was deemed not acceptable to score the project.

3.3 Inter-Rater and Intra-Rater Reliability

All readers were trained to score according to the same scale to ensure accurate, consistent, reliable results. PASC adhered to stringent criteria in its general screening, training, and qualifying procedures as preliminary measures for obtaining high levels of consistency and reliability.

Team leaders conducted “read behinds” as an additional monitoring method. When conducting read behinds, the team leader received student responses and the scores assigned by the reader. The team leader could agree with and confirm the scores, or disagree with the reader’s score and send the paper back for review, citing specific anchor papers as guides.

By default, three percent of each reader's scores appeared in the read-behind application. As more responses were scored, patterns began to emerge and the percentages and types of responses that came through the application were tailored for each scorer.

At least 20 percent of all booklets were read by both the reader and the team leader to check accuracy. An 86 percent overall agreement rate was maintained between readers' scores and team leaders' check scores.

3.4 Calibration Sets

During the scoring process, in addition to scoring the student responses, readers also scored a set of calibration (validity) responses each day. Calibration sets consisted of five student papers of mixed quality in random order, that were pre-scored by expert team leaders who were familiar with the state's scoring parameters. Readers did blind scores of the calibration responses. Readers' scores were compared with known scores and a calibration report prepared. The calibration standard was 80 percent agreement. Any reader who failed to meet the standard was retrained.

3.5 Monitoring Reports

PASC's online scoring system generated many different types of internal monitoring reports that enabled PASC to monitor accuracy of scoring. These reports, computed by individual reader and by team, listed all of a team's readers and provided the results of their scoring on an ongoing basis. Information on these reports included the number of responses read by the readers during the period, the number and percent of invalid responses scored, and the number of responses that received a check score.

The number of responses with check scores provided data for reporting the number of instances of and percent of perfect agreement, the number and percent of responses on which the reader was a point higher or lower than the check scorer, and the number and percent of the responses differing by more than one point (resolution).

The holistic performance by prompt report gave similar information but also included the percentage of responses to which the reader awarded each valid score point. This showed whether the reader tended to distribute scores in a manner similar to other readers. This report was generated daily and cumulatively.

In addition to the reader reports described above, other reports, such as the Project Summary Report, were generated each day to monitor the progress of the orders through the system. This report showed the number and the percent of responses for which first and check-score readings were required and completed.

3.6 Retraining

Room directors conducted group retraining every Monday morning or following any extended break during the project. Individual readers received retraining during the scoring as deemed necessary by the team-leader observations and report results. The need for retraining may be signaled in different ways: high resolution rates resolved against a reader, low or irregular calibration scores, or unsatisfactory perfect-agreement rates or anomalies detected via the read-behind monitoring. Retraining may involve several techniques:

- Discussion of the specific response(s) involved in a resolution of a calibration anomaly
- Discussion of specific papers identified by the read-behind process
- Review of anchors.

The prompt description, i.e., score points for each prompt, their form numbers, IDs, etc., by work groups and grade spans, as well as the rater monitoring summary report are provided in Appendices B.1 and B.2, respectively.

SECTION 4. CLASSICAL ITEM-LEVEL AND MODALITY (SUBTEST) STATISTICS

4.1 Classical Test Theory

There are useful indices available within the framework of classical test theory (CTT) for estimating the precision of the raw test scores and the reliability of assessments. Within CTT, an observed test score is defined as an imprecise estimate of a student's true (and unobservable) ability level and is composed of two components. The first component is referred to as "true score" and is the portion of the observed score that is directly dependent on the student's ability level. The second is an error component (error) and is the portion of the score that is attributable to random error, i.e., the portion of the score attributable to factors unrelated to the student's ability. Error for any student is normally distributed around that student's true score with a mean of zero and an arbitrary standard deviation. Suppose it were possible to give an exam to one student a large number of times without any practice effects. If we were to examine the resulting distribution of scores we would find a normal distribution with a certain mean and a certain standard deviation about the mean. The mean of the resulting distribution is the student's true score according to the definition of error given above. For each student who responds to the exam, error is normally distributed with a mean of zero. However, the standard deviation of the error distribution is idiosyncratic to each student (though it tends to be larger toward the low and high ends of the exam for most tests). If we wanted to estimate what would likely be the standard deviation of this distribution of error for any arbitrary examinee, the best estimate would be the mean of the standard deviations of the error distribution across all examinees. This quantity is called the standard error of measurement (SEM), and is denoted as σ_E . It is defined as:

$$\sigma_E = \sigma_t \sqrt{1 - \rho_t} \quad (5)$$

where σ_t represents the standard deviation of the raw scores for the exam and ρ_t represents the reliability coefficient for the exam.

The standard error of the mean, on the other hand, is an estimate of the magnitude of sampling error associated with the sample mean in the estimation of the population mean. This expected standard mean of sampling errors of the mean is called the standard error of the mean (SE), and is defined as follows:

$$SE = \frac{\sigma}{\sqrt{n}}, \quad (6)$$

where SE represents the standard error of the mean, σ represents the standard deviation of the population, and n represents the number of responses in each sample.

4.2 Item-Level Descriptive Statistics

This section presents the raw-score summary statistics for all items in the 2006 MI-ELPA within the framework of CTT. The p -value for each item is defined as the proportion of students who answer an item correctly for the multiple-choice items. A high p -value means that an item is easy; a low p -value means that an item is difficult. For the constructed-response items, the p -value is reported as the average number of points out of the maximum number of possible points.

The point biserial correlation for each item is an index of the association between the item score and the total-test score. It shows the ability of the item to discriminate between low-ability and high-ability students. An item with a high point biserial correlation discriminates more effectively between the low- and the high-ability students than an item with a low point biserial correlation.

The item-level statistics for the operational and the embedded field-test items for the 2006 MI-ELPA are presented in Appendices C.1–C.4 by grade span (level) and form. The tables are grouped by modalities, i.e., Listening, Reading, Writing, and Speaking. The following item information and statistics are presented for each item:

- Item number based on the items' sequential appearance in the form by modality
- Item type (multiple-choice or constructed-response by score point indication, e.g., C2, C3)
- Item designation as core (C) or embedded field-test (FT)
- Maximum number of possible points
- N-Count (number of students)
- Response options for multiple-choice items and percentage of students obtaining each score point for constructed-response items
- Omits (percentage of students omitting an item)
- p -value for multiple-choice items (percentage of examinees who answered the item correctly) and item mean for constructed-response items (average number of points earned out of the maximum number of possible points)
- Point Biserial/Item-to-Total Correlation (index of discrimination between high- and low-scoring students)

4.3 Measure of Central Tendency

The classical measures of central tendency for the MI-ELPA scores are also reported by the Listening, Speaking, Reading, Writing, and Comprehension modalities. The Comprehension modality, however, consists of items selected from the Listening and Reading components of the MI-ELPA.

The classical measures of central tendency, variability, and score precision are presented in Table 4.1 by grade span for each modality as well as for the total test. Table 4.2 presents the same statistics by grade. The tables include the following:

2006 MI-ELPA Technical Manual

- Number of items
- Maximum score attainable
- N-Count (sample size)
- RS Mean (average raw score)
- SD (standard deviation of raw scores)
- SE (standard error of the mean)

Table 4.1: Summary Statistics of MI-ELPA Modalities by Grade Span

Grade Span	Modality/Test	Number of Items	Max Points	N-Count	RS Mean	SD	SE
K-2	Listening	21	21	22085	14.07	3.21	0.02
	Speaking	13	23	22085	17.81	4.55	0.03
	Reading	22	22	22085	13.57	4.73	0.03
	Writing	13	19	22085	10.32	5.84	0.04
	Comprehension	33	33	22085	20.65	5.44	0.04
	Total Test	69	85	22085	55.77	15.76	0.11
3-5	Listening	21	21	16784	16.92	3.20	0.02
	Speaking	12	23	16784	19.94	3.65	0.03
	Reading	21	21	16784	13.42	4.10	0.03
	Writing	18	23	16784	17.94	3.74	0.03
	Comprehension	36	36	16784	25.61	5.83	0.05
	Total Test	72	88	16784	68.23	12.31	0.09
6-8	Listening	21	21	12640	15.98	3.16	0.03
	Speaking	13	25	12640	21.66	4.37	0.04
	Reading	23	23	12640	15.10	4.32	0.04
	Writing	17	23	12640	16.97	3.89	0.03
	Comprehension	32	32	12640	22.00	5.13	0.05
	Total Test	74	92	12640	69.72	13.33	0.12
9-12	Listening	24	24	10344	17.97	4.22	0.04
	Speaking	13	25	10344	21.24	4.46	0.04
	Reading	25	25	10344	17.36	5.14	0.05
	Writing	18	25	10344	17.03	4.51	0.04
	Comprehension	33	33	10344	23.56	5.60	0.06
	Total Test	80	99	10344	73.60	15.86	0.16

Note. 1. Total Test does not include the Comprehension modality.
 2. The total n-counts for grade spans were obtained after deleting all raw scores of 999 while “Omits” and “Blanks” were scored as “0s.”

Table 4.2: Summary Statistics of MI-ELPA Modalities by Grade

Grade	Modality/Test	Number of Items	Max Points	N Count	RS Mean	SD	SE
K	Listening	21	21	7773	12.00	2.80	0.03
	Speaking	13	23	7773	15.34	4.81	0.05
	Reading	22	22	7773	9.48	3.31	0.04
	Writing	13	19	7773	4.60	3.74	0.04
	Comprehension	33	33	7773	16.31	3.90	0.04
	Total Test	69	85	7773	41.42	10.91	0.12
1	Listening	21	21	7507	14.29	2.72	0.03
	Speaking	13	23	7507	18.44	3.94	0.05
	Reading	22	22	7507	14.32	3.49	0.04
	Writing	13	19	7507	11.97	4.24	0.05
	Comprehension	33	33	7507	21.22	4.18	0.05
	Total Test	69	85	7507	59.02	11.45	0.13
2	Listening	21	21	6805	16.19	2.62	0.03
	Speaking	13	23	6805	19.94	3.42	0.04
	Reading	22	22	6805	17.42	3.47	0.04
	Writing	13	19	6805	15.03	3.52	0.04
	Comprehension	33	33	6805	24.98	4.34	0.05
	Total Test	69	85	6805	68.58	10.70	0.13
3	Listening	21	21	6116	16.17	3.31	0.04
	Speaking	12	23	6116	19.47	3.80	0.05
	Reading	21	21	6116	11.97	3.94	0.05
	Writing	18	23	6116	16.94	3.90	0.05
	Comprehension	36	36	6116	23.66	5.67	0.07
	Total Test	72	88	6116	64.55	12.32	0.16
4	Listening	21	21	5468	17.03	3.14	0.04
	Speaking	12	23	5468	20.05	3.60	0.05
	Reading	21	21	5468	13.65	3.93	0.05
	Writing	18	23	5468	18.14	3.63	0.05
	Comprehension	36	36	5468	25.90	5.66	0.08
	Total Test	72	88	5468	68.87	11.93	0.16
5	Listening	21	21	5200	17.69	2.92	0.04
	Speaking	12	23	5200	20.38	3.45	0.05
	Reading	21	21	5200	14.89	3.87	0.05
	Writing	18	23	5200	18.92	3.35	0.05
	Comprehension	36	36	5200	27.59	5.46	0.08
	Total Test	72	88	5200	71.88	11.45	0.16
6	Listening	21	21	4646	15.68	3.15	0.05
	Speaking	13	25	4646	21.58	4.32	0.06
	Reading	23	23	4646	14.49	4.27	0.06
	Writing	17	23	4646	16.56	3.90	0.06
	Comprehension	32	32	4646	21.33	5.06	0.07
	Total Test	74	92	4646	68.31	13.12	0.19

Table 4.2: Summary Statistics of MI-ELPA Modalities by Grade (Continued)

Grade	Modality/Test	Number of Items	Max Points	N Count	RS Mean	SD	SE
7	Listening	21	21	4164	16.01	3.13	0.05
	Speaking	13	25	4164	21.59	4.44	0.07
	Reading	23	23	4164	15.10	4.30	0.07
	Writing	17	23	4164	16.99	3.87	0.06
	Comprehension	32	32	4164	22.01	5.06	0.08
	Total Test	74	92	4164	69.69	13.30	0.21
8	Listening	21	21	3830	16.32	3.18	0.05
	Speaking	13	25	3830	21.85	4.34	0.07
	Reading	23	23	3830	15.85	4.28	0.07
	Writing	17	23	3830	17.45	3.83	0.06
	Comprehension	32	32	3830	22.81	5.15	0.08
	Total Test	74	92	3830	71.47	13.41	0.22
9	Listening	24	24	3967	17.33	4.42	0.07
	Speaking	13	25	3967	20.69	5.11	0.08
	Reading	25	25	3967	16.43	5.33	0.08
	Writing	18	25	3967	16.19	4.86	0.08
	Comprehension	33	33	3967	22.60	5.82	0.09
	Total Test	80	99	3967	70.65	17.20	0.27
10	Listening	24	24	2899	18.05	4.20	0.08
	Speaking	13	25	2899	21.38	4.18	0.08
	Reading	25	25	2899	17.54	5.05	0.09
	Writing	18	25	2899	17.2	4.32	0.08
	Comprehension	33	33	2899	23.7	5.54	0.1
	Total Test	80	99	2899	74.16	15.35	0.29
11	Listening	24	24	1984	18.44	3.92	0.09
	Speaking	13	25	1984	21.61	3.90	0.09
	Reading	25	25	1984	18.03	4.78	0.11
	Writing	18	25	1984	17.68	4.05	0.09
	Comprehension	33	33	1984	24.26	5.20	0.12
	Total Test	80	99	1984	75.75	14.14	0.32
12	Listening	24	24	1494	18.89	3.84	0.10
	Speaking	13	25	1494	21.95	3.57	0.09
	Reading	25	25	1494	18.57	4.80	0.12
	Writing	18	25	1494	18.06	4.08	0.11
	Comprehension	33	33	1494	24.87	5.21	0.13
	Total Test	80	99	1494	77.47	13.77	0.36

Note. 1. Total Test does not include the Comprehension modality.
 2. The total n-count for each grade was obtained after deleting all raw scores of 999 while “Omits” and “Blanks” were scored as “0s.”

4.4 MI-ELPA Stand-Alone Field-Test Item Statistics

For the 2006 field testing for the MI-ELPA, the grade spans were broken down as KIN (Grade 1), PRI (Grades 2–3), ELE (Grades 4–6), MID (Grades 7–9), and HGH (Grade 10–12). The field-testing grade spans were different from the operational grade spans because field testing was completed in fall 2006 while operational testing will be administered in spring 2007. Because of the time span between field testing and operational testing, all kindergarten students would be in Grade 1 and should be tested at that grade. A similar pattern was followed by the rest of the grades. Multiple forms were administered for each grade span (KIN: 7 forms; PRI: 8 forms; ELE: 8 forms, MID: 6 forms; HGH: 6 forms). These field-test items were independently administered, i.e., they were not embedded within operational forms.

The targeted n-counts could not be met because this was not mandatory testing and some schools did not participate in spite of their commitment. A second sampling was requested by OEAA when the first sampling did not provide the required n-counts. However, due to the time constraints, the data were not received in time for the selection process. Therefore, in the field-test analysis, the n-counts were far below expectations. The n-counts with the respective classical item statistics by modality, grade span, and form are provided in Appendix C.5, while the same information is provided for each grade in Appendix C.6.

SECTION 5. RELIABILITY

Reliability is the degree to which scores remain consistent over an assessment procedure (Nitko, 2004). Further defined, reliability is the degree to which students' assessment results are consistent when a) they complete the same task on two or more occasions; b) two or more raters evaluate their performance on the same task; or c) they complete two or more parallel tasks on one or more occasions. Consistency of scores over repeated assessment and/or with different raters is the underlying feature of reliability.

5.1 Internal Consistency Reliability

The internal consistency of a test investigates the stability of scores from one sample of content to another. Several methods can be used to estimate the internal consistency of a test. One approach is to split all test questions into two groups and then correlate student scores on the two half-tests. This is known as a split-half estimate of reliability. This method avoids the implications of any changes in the individual by administering only a single test. If scores have a high rate of correlation on the two half-tests, it can be concluded that the test questions complement one another, function well as a group, and measure similar concepts. This also suggests that measurement error is minimal.

The split-half method's decision about which questions contribute to each half-test's score can have an impact on the resulting correlation. Harcourt uses Cronbach's coefficient alpha statistic (Cronbach, 1951) to avoid this concern about the split-half method. The coefficient alpha is the average split-half correlation based on all possible divisions of a test into two parts. The coefficient alpha can be used to estimate the internal consistency of both dichotomously (right or wrong, 0 or 1 score values) and polytomously (a wide range of score values) scored test items. Coefficient alpha is computed by the following formula:

$$\alpha = \frac{I}{I-1} \left(1 - \frac{\sum_i s_i^2}{S_x^2} \right), \quad (7)$$

where I represents the number of items on the test, s_i^2 represents the variance of item i , and S_x^2 represents the total test variance.

5.2 Classical SEM (based on Classical Test Theory)

Since no assessment measures ability with perfect consistency, it is useful to take into account the likely size of measurement errors. One way to describe the inconsistency of assessment results is to assess a student on multiple occasions and note how much the scores vary. Repeatedly measuring a student can only be done hypothetically, however, but if you could assess a student on multiple occasions you would obtain a collection of the student's obtained scores. The scores would cluster around an average value. The standard deviation, or spread, of these obtained scores is known as the standard error of measurement (SEM).

The SEM is another index of reliability and provides an estimate of the amount of error in an individual's observed test score. The individual's observed total score is considered the estimate of the person's true score. Because the standard error of measurement is inversely related to the reliability of a test, the greater the reliability, the less the standard error of measurement, and the more confidence one may have in the accuracy, or precision, of the observed test score. The measurement error is commonly expressed in terms of standard deviation units; that is, the standard error of measurement is the standard deviation of the measurement error distribution. The standard error of measurement is calculated with the following equation:

$$SEM = SD\sqrt{1-r_{xx}} \Leftrightarrow s_e = s_x\sqrt{1-\frac{s_t^2}{s_x^2}}, \quad (8)$$

where $SEM (=s_e)$ refers to the standard error of measurement, $SD (=s_x)$ represents the standard deviation unit of the scale for a test, r_{xx} represents the reliability coefficient for a sample test (or estimate of ρ_{xx} , which is a population reliability coefficient), s_t^2 represents the estimate of σ_T^2 , and s_x^2 represents the estimate of σ_X^2 .

5.3 Conditional SEM (based on Item Response Theory)

Unlike the SEM based on the CTT, the SEM based on the item response theory (IRT) is not the same for all persons. For example, if a person answers either a few items or a large number of items correctly (extreme score), the SEM is greater than if the person answers a moderate number of items correctly. This implies that the SEM depends on the total score (Andrich & Luo, 2004).

Under the Rasch model, the SEM for each person is as follows:

$$\sigma_{\hat{\beta}} = \frac{1}{\sqrt{\sum_{i=1}^L p_{vi}(1-p_{vi})}}, \quad (9)$$

where v is subscript for a person, i is subscript for an item, L represents length of the test, $\hat{\beta}$ represents ability estimate, and p_{vi} represents the probability that a person answers an item correctly and is defined as follows:

$$p_{vi} = \frac{e^{\beta_v - \delta_i}}{1 + e^{\beta_v - \delta_i}}, \quad (10)$$

where β_v represents person v 's ability and δ_i represents the item's difficulty.

A confidence band can be used in interpreting the ability estimate. For example, an approximate 68 percent confidence interval for $\hat{\beta}$ is given by $\hat{\beta} \pm SEM$.

Note that the SEM for item difficulty is smallest when the probability of passing is close to the probability of failing. That is, when an item is near the threshold level for many persons in the sample, the SEM is small (Embretson & Reise, 2000).

The conditional SEMs are presented in the raw score to scale score conversion tables in Appendix D.

5.4 Inter-Rater Reliability

Another source of measurement error lies in the evaluation of student work. Inter-rater reliability, as explained in Section 3 of this manual, investigates the extent to which examinees would obtain the same score if the assessment task is scored two or more times by the same rater or different raters. One way to estimate this type of reliability is to have two raters score each student's paper and then obtain the correlation. In this case, consistency is defined as similarity of students' rank orderings by two raters. Another way to obtain evidence of inter-rater reliability is to calculate the percent agreement between raters. If raters always agree in their assignment of scores, there is 100 percent agreement. If raters never agree in their assignment of scores, there is 0 percent agreement. The choice between using a correlation coefficient or percent agreement depends upon whether students' absolute (actual) or relative (rank order) score level is important for a particular interpretation and use. See Appendices B.1 and B.2 for the scoring prompt specification and the results of the analyses of rater agreement for MI-ELPA writing prompts.

5.5 Reliability of Each of the Five Modalities

Table 5.1 provides the raw-score descriptive statistics and reliabilities by grade and modalities. It includes the following information:

- Number of items
- Maximum number of possible points
- Number of students (N-Count)
- Means and standard deviations in raw scores (RS Mean; SD)
- Cronbach's alpha internal consistency reliability
- Standard error of measurement (SEM)

Table 5.1: Descriptive Statistics and Reliability by Grade and Modality

Grade	Modality/Test	Number of Items	Max Points	N-Count	RS Mean	SD	Reliability	SEM
K	Listening	21	21	7773	12.00	2.80	0.55	1.88
	Speaking	13	23	7773	15.34	4.81	0.93	1.27
	Reading	22	22	7773	9.48	3.31	0.64	1.99
	Writing	13	19	7773	4.60	3.74	0.83	1.54
	Comprehension	33	33	7773	16.31	3.90	0.61	2.44
	Total Test	69	85	7773	41.42	10.91	0.89	3.62
1	Listening	21	21	7507	14.29	2.72	0.59	1.74
	Speaking	13	23	7507	18.44	3.94	0.91	1.18
	Reading	22	22	7507	14.32	3.49	0.75	1.75
	Writing	13	19	7507	11.97	4.24	0.81	1.85
	Comprehension	33	33	7507	21.22	4.18	0.71	2.25
	Total Test	69	85	7507	59.02	11.45	0.91	3.44
2	Listening	21	21	6805	16.19	2.62	0.67	1.51
	Speaking	13	23	6805	19.94	3.42	0.89	1.13
	Reading	22	22	6805	17.42	3.47	0.83	1.43
	Writing	13	19	6805	15.03	3.52	0.72	1.86
	Comprehension	33	33	6805	24.98	4.34	0.79	1.99
	Total Test	69	85	6805	68.58	10.70	0.92	3.03
3	Listening	21	21	6116	16.17	3.31	0.75	1.66
	Speaking	12	23	6116	19.47	3.80	0.87	1.37
	Reading	21	21	6116	11.97	3.94	0.75	1.97
	Writing	18	23	6116	16.94	3.90	0.76	1.91
	Comprehension	36	36	6116	23.66	5.67	0.80	2.54
	Total Test	72	88	6116	64.55	12.32	0.92	3.48
4	Listening	21	21	5468	17.03	3.14	0.75	1.57
	Speaking	12	23	5468	20.05	3.60	0.84	1.44
	Reading	21	21	5468	13.65	3.93	0.78	1.84
	Writing	18	23	5468	18.14	3.63	0.74	1.85
	Comprehension	36	36	5468	25.90	5.66	0.82	2.40
	Total Test	72	88	5468	68.87	11.93	0.92	3.37
5	Listening	21	21	5200	17.69	2.92	0.75	1.46
	Speaking	12	23	5200	20.38	3.45	0.81	1.50
	Reading	21	21	5200	14.89	3.87	0.80	1.73
	Writing	18	23	5200	18.92	3.35	0.71	1.80
	Comprehension	36	36	5200	27.59	5.46	0.83	2.25
	Total Test	72	88	5200	71.88	11.45	0.92	3.24
6	Listening	21	21	4646	15.68	3.15	0.76	1.54
	Speaking	13	25	4646	21.58	4.32	0.90	1.37
	Reading	23	23	4646	14.49	4.27	0.79	1.96
	Writing	17	23	4646	16.56	3.90	0.87	1.41
	Comprehension	32	32	4646	21.33	5.06	0.80	2.26
	Total Test	74	92	4646	68.31	13.12	0.94	3.21

Table 5.1: Descriptive Statistics and Reliability by Grade and Modality (Continued)

Grade	Modality/Test	Number of Items	Max Points	N-Count	RS Mean	SD	Reliability	SEM
7	Listening	21	21	4164	16.01	3.13	0.78	1.47
	Speaking	13	25	4164	21.59	4.44	0.91	1.33
	Reading	23	23	4164	15.10	4.30	0.80	1.92
	Writing	17	23	4164	16.99	3.87	0.88	1.34
	Comprehension	32	32	4164	22.01	5.06	0.81	2.21
	Total Test	74	92	4164	69.69	13.30	0.94	3.26
8	Listening	21	21	3830	16.32	3.18	0.80	1.42
	Speaking	13	25	3830	21.85	4.34	0.91	1.30
	Reading	23	23	3830	15.85	4.28	0.82	1.82
	Writing	17	23	3830	17.45	3.83	0.89	1.27
	Comprehension	32	32	3830	22.81	5.15	0.83	2.12
	Total Test	74	92	3830	71.47	13.41	0.95	3.00
9	Listening	24	24	3967	17.33	4.42	0.83	1.82
	Speaking	13	25	3967	20.69	5.11	0.98	0.72
	Reading	25	25	3967	16.43	5.33	0.87	1.92
	Writing	18	25	3967	16.19	4.86	0.84	1.94
	Comprehension	33	33	3967	22.60	5.82	0.84	2.33
	Total Test	80	99	3967	70.65	17.20	0.96	3.44
10	Listening	24	24	2899	18.05	4.20	0.83	1.73
	Speaking	13	25	2899	21.38	4.18	0.95	0.93
	Reading	25	25	2899	17.54	5.05	0.87	1.82
	Writing	18	25	2899	17.20	4.32	0.80	1.93
	Comprehension	33	33	2899	23.70	5.54	0.83	2.28
	Total Test	80	99	2899	74.16	15.35	0.95	3.43
11	Listening	24	24	1984	18.44	3.92	0.81	1.71
	Speaking	13	25	1984	21.61	3.90	0.93	1.03
	Reading	25	25	1984	18.03	4.78	0.86	1.79
	Writing	18	25	1984	17.68	4.05	0.78	1.90
	Comprehension	33	33	1984	24.26	5.20	0.82	2.21
	Total Test	80	99	1984	75.75	14.14	0.94	3.46
12	Listening	24	24	1494	18.89	3.84	0.81	1.67
	Speaking	13	25	1494	21.95	3.57	0.91	1.07
	Reading	25	25	1494	18.57	4.80	0.87	1.73
	Writing	18	25	1494	18.06	4.08	0.79	1.87
	Comprehension	33	33	1494	24.87	5.21	0.83	2.15
	Total Test	80	99	1494	77.47	13.77	0.94	3.37

Note. 1. Total Test does not include the Comprehension modality.

2. The total n-counts for each grade was obtained after deleting all raw scores of 999 while “Omits” and “Blanks” were scored as “0s.”

5.6 Reliability of Classification Decision at Proficient Cut

Based on the MI-ELPA scale scores, student performance is classified into one of four proficiency levels. While it is always important to know the reliability of student scores in any examination, it is of even greater importance to assess the reliability of the decisions based on these scores. Evaluation of the reliability of classification decisions is performed through estimation of the probabilities of correct and consistent classification of student performance. Procedures from Livingston and Lewis (1995) were applied to derive measures of the accuracy and consistency of the classifications. Brief descriptions of the procedures used and results obtained are presented in this section.

The accuracy of decisions is the extent to which decisions would agree with those that would be made if each student could somehow be tested with all possible forms of the examination. The consistency of decisions is the extent to which decisions would agree with the decisions that would have been made if the students had taken a parallel form of the MI-ELPA, equal in difficulty and covering the same content as the form they actually took. These ideas are shown schematically in Figures 5.1 and 5.2. Please note that the term *Achieves Proficient Status* refers to the proficient category on the Listening/Speaking and Reading/Writing combinations score, and *Does Not Achieve Proficient Status* refers to all categories below proficient status.

		Decision made on a form actually taken	
		<i>Does Not Achieve Proficient Status</i>	<i>Achieves Proficient Status</i>
True status made on all-forms average	<i>Does Not Achieve Proficient Status</i>	Correct Classification	Misclassification
	<i>Achieves Proficient Status</i>	Misclassification	Correct Classification

Figure 5.1: Classification Accuracy

		Decision made on the 2nd form taken	
		<i>Does Not Achieve Proficient Status</i>	<i>Achieves Proficient Status</i>
Decision made on the 1st form taken	<i>Does Not Achieve Proficient Status</i>	Correct Classification	Misclassification
	<i>Achieves Proficient Status</i>	Misclassification	Correct Classification

Figure 5.2: Classification Consistency

Note. Figures 5.1 and 5.2 are adapted from Young and Yoon (1998).

In Figure 5.1, accurate classifications occur when the decision made on the basis of the all-forms average (or true score) agrees with the decision made on the basis of the form actually taken.

Misclassifications occur when, for example, a student who actually accomplished *Does Not Achieve Proficient Status* on the basis of his or her all-forms average is classified incorrectly as accomplishing *Achieves Proficient Status*. Consistent classification occurs (Figure 5.2) when two forms agree on the classification of a student as either *Achieves Proficient Status* or *Does Not Achieve Proficient Status*, whereas inconsistent classification occurs when the decisions made by the forms differ.

These analyses make use of the techniques outlined and implemented by Hanson (1991), Haertel (1996), Livingston and Lewis (1995), and Young and Yoon (1998). The software developed by Hanson (1995) was used for the analyses. Estimates of decision accuracy and consistency were made for the *Achieves Proficient Status* cut-scores on the Listening/Speaking and Reading/Writing scores reported in the MI-ELPA.

Table 5.2 presents the results of the decision accuracy and consistency of the Achieves Proficient Status cut scores for the Listening/Speaking and Reading/Writing scores. The table includes the proportions of False Positive and False Negative classifications. The sum of values of Accuracy, False Positive, and False Negative is equal to 1.00, but due to rounding the table values may or may not equal 1.00. False Positive and False Negative classifications refer to the mismatch between student true scores and observed scores. The False Positive value is the proportion of student scores misclassified to the category *Achieves Proficient Status* when student scores do not meet proficient status. The False Negative value is the proportion of student scores misclassified to the category *Does Not Achieve Proficient Status* when student scores actually do meet proficient status. Table 5.2 contains the following:

- Consistent classifications
- Accurate classifications
- False positives
- False negatives

The table illustrates the general rule that decision consistency will be less than decision accuracy. It should also be noted that the students who achieved proficient status for the Listening/Speaking combination ranged from 0.76 to 0.89 and the students who achieved proficient status for the Reading/Writing combination ranged from 0.88 to 0.99.

Table 5.2: Decision and Consistency Table by Grade

Grade	Test	Accuracy	False Positives	False Negatives	Consistency
K	Total MI-ELPA	0.89	0.06	0.05	0.85
1	Total MI-ELPA	0.93	0.03	0.04	0.90
2	Total MI-ELPA	0.96	0.02	0.03	0.94
3	Total MI-ELPA	0.96	0.02	0.02	0.95
4	Total MI-ELPA	0.97	0.01	0.02	0.96
5	Total MI-ELPA	0.97	0.01	0.01	0.96
6	Total MI-ELPA	0.96	0.02	0.02	0.94
7	Total MI-ELPA	0.95	0.02	0.03	0.93
8	Total MI-ELPA	0.96	0.02	0.02	0.94
9	Total MI-ELPA	0.95	0.03	0.02	0.93
10	Total MI-ELPA	0.95	0.02	0.03	0.93
11	Total MI-ELPA	0.93	0.03	0.04	0.91
12	Total MI-ELPA	0.93	0.03	0.04	0.90

Note. The sum of Accuracy, False Positives, and False Negatives may not add up to 1.00 because of rounding.

SECTION 6. VALIDITY

For the 2006 administration of the MI-ELPA, Harcourt’s ELL item bank was used to construct one form per grade span. Besides the Harcourt ELL item bank, MWAC was also used for item procurement (See Appendix A for the 2006 test blueprints). Special calibration studies were conducted on all items in the Harcourt ELL item bank in order to obtain both traditional and Rasch item statistics¹. A wealth of item information was gathered through these calibration studies. Among the statistics included are *p*-values, point-biserials, Rasch difficulty, and standard error of the Rasch difficulty. In addition to the item statistics, several intact forms have been created. Assessments constructed from the Harcourt ELL item bank support the validity-related standards set forth in the *Standards for Educational and Psychological Testing*. Our judgments about test validity are based on the following sources of evidence of validity²:

- Test content—“an analysis of the relationship between a test’s content and the construct it is intended to measure” (p. 11)
- Internal structure—“the degree to which the relationships among test items and test components conform to the construct on which the proposed test score interpretations are made” (p. 13)
- Relationships to other variables—“analyses of the relationship of test scores to variables external to the test” (p. 13)

6.1 Test Content

Evidence of validity based on test content is revealed by the extent to which the material on the test represents an appropriate sampling of skills, knowledge, and understanding of the domain tested. As part of the development of the Harcourt ELL item bank, item writers were trained to write items representative of the intent of the instructional standards set forth in the test blueprint. In addition, a critical part of the item review process included the appropriateness of the match of the item to the instructional standard being assessed. Only those items relating specifically to an instructional standard (refer to the following URL: http://www.michigan.gov/mde/0,1607,7-140-22709_40192---,00.html for the Michigan Learning Standards) were included in the test forms.

6.2 Evidence of the Test Content for the MI-ELPA

In order for the 2006 MI-ELPA to accurately measure the Michigan Learning Standards, the items in the Harcourt ELL item bank were reviewed to match the standards for each grade span. The item mapping provided in the blueprints together with the matching of items to a particular Michigan Learning Standard for creating the 2006 MI-ELPA gave concrete evidence for the alignment to the Michigan Learning Standards.

¹For details of the features of item bank including research studies, please refer to the Stanford English Language Proficiency Test Technical Manual, 2005, Harcourt Assessment, Inc.

²The page number in the parentheses is the page number in the *Standards for Educational and Psychological Testing*, 1999.

6.3 Internal Structure

Because an English language proficiency test should be able to detect performance and proficiency differences among students, it is important to examine how well each item functions consistently with the overall intent of the test. Biserial correlation coefficients reveal how well an item discriminates between high- and low-achieving students. In developing test forms, we examined the fit between the construct being assessed in terms of the way it was assessed and the way students were able to respond. Content experts were asked to examine the test blueprints and items to be sure that the test would logically relate to the most current empirical and theoretical understanding of the constructs being assessed.

6.4 Evidence of the Internal Structure of the MI-ELPA

An assessment procedure should not be a random collection of assessment tasks or test questions. Each task in the assessment should contribute positively to the total result. The interrelationship among the tasks on an assessment is known as the internal structure of the assessment. Typical questions that investigate the relationships among assessment parts include (Nitko, 2004):

- Do all of the assessment tasks “work together” so that each task contributes positively toward assessing the quality of interest?
- If different parts of the assessment procedure are to provide unique information, do the results support this uniqueness?
- If different parts of the assessment procedure are to provide the same or similar information, do the results support this?

In order to investigate the answers to these questions, correlations were obtained between the four modalities. Table 6.1 presents the intercorrelations among the four modalities by grade. The evidence of internal structure of the 2006 MI-ELPA is also depicted by the point biserial correlation coefficient and fit statistics. Appendices C.1– C.6 and F.1– F.4 (IRT Statistics) provide these statistics for the 2006 MI-ELPA.

Table 6.1: Intercorrelations Among Modalities by Grade

Grade	Modality/Test	Correlation Coefficient					
		Listening	Speaking	Reading	Writing	Comprehension	Total Test
K	Listening	1.00					
	Speaking	0.38	1.00				
	Reading	0.41	0.30	1.00			
	Writing	0.38	0.35	0.60	1.00		
	Comprehension	0.83	0.38	0.77	0.52	1.00	
	Total Test	0.68	0.75	0.75	0.78	0.79	1.00
1	Listening	1.00					
	Speaking	0.42	1.00				
	Reading	0.51	0.41	1.00			
	Writing	0.47	0.48	0.71	1.00		
	Comprehension	0.85	0.45	0.84	0.63	1.00	
	Total Test	0.71	0.75	0.83	0.86	0.85	1.00
2	Listening	1.00					
	Speaking	0.47	1.00				
	Reading	0.60	0.50	1.00			
	Writing	0.52	0.53	0.73	1.00		
	Comprehension	0.87	0.53	0.88	0.67	1.00	
	Total Test	0.76	0.77	0.87	0.86	0.89	1.00
3	Listening	1.00					
	Speaking	0.58	1.00				
	Reading	0.59	0.44	1.00			
	Writing	0.62	0.56	0.65	1.00		
	Comprehension	0.86	0.56	0.90	0.69	1.00	
	Total Test	0.83	0.78	0.82	0.86	0.91	1.00
4	Listening	1.00					
	Speaking	0.61	1.00				
	Reading	0.62	0.47	1.00			
	Writing	0.64	0.58	0.66	1.00		
	Comprehension	0.86	0.59	0.92	0.70	1.00	
	Total Test	0.85	0.79	0.84	0.86	0.92	1.00
5	Listening	1.00					
	Speaking	0.63	1.00				
	Reading	0.63	0.51	1.00			
	Writing	0.65	0.61	0.65	1.00		
	Comprehension	0.85	0.61	0.93	0.71	1.00	
	Total Test	0.85	0.81	0.84	0.86	0.92	1.00
6	Listening	1.00					
	Speaking	0.57	1.00				
	Reading	0.64	0.45	1.00			
	Writing	0.68	0.61	0.70	1.00		
	Comprehension	0.88	0.54	0.89	0.73	1.00	
	Total Test	0.84	0.80	0.84	0.89	0.89	1.00
7	Listening	1.00					
	Speaking	0.59	1.00				
	Reading	0.65	0.48	1.00			
	Writing	0.68	0.64	0.70	1.00		
	Comprehension	0.88	0.55	0.90	0.73	1.00	
	Total Test	0.84	0.81	0.84	0.89	0.89	1.00

Table 6.1: Intercorrelations Among Modalities by Grade (Continued)

8	Listening	1.00					
	Speaking	0.63	1.00				
	Reading	0.66	0.52	1.00			
	Writing	0.70	0.67	0.72	1.00		
	Comprehension	0.89	0.60	0.90	0.75	1.00	
	Total Test	0.85	0.83	0.85	0.90	0.91	1.00
9	Listening	1.00					
	Speaking	0.62	1.00				
	Reading	0.75	0.58	1.00			
	Writing	0.72	0.66	0.76	1.00		
	Comprehension	0.93	0.62	0.90	0.76	1.00	
	Total Test	0.88	0.82	0.89	0.90	0.92	1.00
10	Listening	1.00					
	Speaking	0.59	1.00				
	Reading	0.74	0.58	1.00			
	Writing	0.70	0.61	0.75	1.00		
	Comprehension	0.92	0.59	0.90	0.75	1.00	
	Total Test	0.87	0.80	0.90	0.89	0.92	1.00
11	Listening	1.00					
	Speaking	0.57	1.00				
	Reading	0.72	0.52	1.00			
	Writing	0.66	0.58	0.70	1.00		
	Comprehension	0.91	0.55	0.90	0.70	1.00	
	Total Test	0.87	0.78	0.88	0.87	0.91	1.00
12	Listening	1.00					
	Speaking	0.55	1.00				
	Reading	0.70	0.52	1.00			
	Writing	0.66	0.54	0.70	1.00		
	Comprehension	0.90	0.54	0.89	0.71	1.00	
	Total Test	0.86	0.75	0.89	0.86	0.92	1.00

Note. Total Test does not include the Comprehension modality.

To help interpret Table 6.1, Harcourt Content Development experts and psychometricians explored the existing research from Educational Testing Service (ETS), followed by some explanation of Table 6.1.

Research of intercorrelations of language proficiency assessment subtests for young adults

- Listening and Reading are highly correlated: .69 for TOEFL Listening/Reading (Educational Testing Service 1997) and .84 for SLEP Listening/Reading (Educational Testing Service 1991)
- Reading and Writing are moderately correlated: .56–.59 for TOEFL Reading/Test of Written English (Educational Testing Service 1996)
- Historically, the language domain pairs of Listening and Speaking and Reading and Writing are moderately to highly correlated while Speaking and Writing are not correlated.

Kindergarten

- Students in this age group do not usually read or write yet, but they can have Listening and Speaking skills.
- The expected outcome is that neither Reading nor Writing will correlate with Listening or Speaking.

Grades 1–8

- A steady increase in the correlation between Writing and Speaking is observed.
- A possible explanation is that, in general, students during this age span experience expanding use of and development in their Writing skills. At the same time, demands on the Listening skills of this age group remain fairly static with only moderate development.

Grades 9–12

- A steady decrease in the correlation between Writing and Speaking is observed.
- A possible explanation is that by high school, there is an increased focus on use of Writing skills, especially an increased focus on academic content. Requirements of high school age student Listening skills also decrease, but not nearly at as steep a curve as Writing.

Similar arguments may be made for the correlational behavior between Listening and Writing in grades 1–12.

6.5 Relationships to Other Variables

For the items in the Harcourt ELL item bank, evidence of validity based on relationships to other variables is revealed by examining the following studies. Since the 2006 administration of the MI-ELPA was partly based on the Harcourt ELL item bank, the evidence of validity is reported on the SELP.

Performance Differences Between Native and Non-Native English-Speaking Students Taking the SELP

The major purpose of this study was to compare scores achieved by native and non-native speakers of English on three SELP multiple-choice subtests. The mean scores obtained on the Listening, Writing Conventions, and Reading subtests were used to identify any group differences between native and non-native speakers of English. The results of this study indicate that there is a significant difference in the scores between the non-native and native English-speaking students. As expected, the native speakers scored higher than the non-native speakers. The analysis of variance results support this expectation.

Relationship between the SELP and the Stanford Diagnostic Reading Test (SDRT)

The results of this study support the hypothesis that there is a high positive correlation between SELP and SDRT. The Pearson Product-Moment correlation coefficients range from 0.76 to 0.80. The data reveal that students who scored high on the SELP also scored high on the SDRT; similarly, students who scored low on the SELP also scored low on the SDRT.

Relationship between the SELP and the Abbreviated Reading Subtest of the Stanford Achievement Test Series, Ninth Edition (Stanford 9)

The analyses for this study are grouped by Stanford 9 test levels. The results of this study show that there is a low positive correlation between scores earned on the SELP multiple-choice subtests and the Stanford 9 Abbreviated Reading subtest. The Pearson Product-Moment correlation coefficient ranges from 0.33 to 0.53. The correlations show that high scores on the SELP correspond with high scores on the Abbreviated Stanford 9 Reading subtest. Similarly, low scores on the SELP correspond with low scores on the Abbreviated Stanford 9 Reading subtest.

SECTION 7. CALIBRATION, EQUATING, AND SCALING

The items on the MI-ELPA were analyzed within the framework of Item Response Theory (IRT). IRT is widely used because of the advantages it confers upon the exam consumers. It promotes equity of results from year to year, through what has been referred to as test-free measurement. Simply stated, test-free measurement means that, given a student's responses to two exams scale using IRT, that student will achieve the same scale score on both exams except for measurement error. This holds true regardless of differences in the overall difficulties of the exams. In other words, measurement is test-free in the sense that the results are dependent only upon the ability of the student and are independent of the item difficulties.

The Rasch model (Rasch, 1960) for dichotomous items and the Partial Credit Model (PCM) (Masters, 1982) for polytomous items were used to develop, calibrate, equate, and scale the MI-ELPA. These measurement models are regularly used to construct test forms, for scaling and equating, and to develop and maintain large item banks. All item and test analyses, including item-fit analysis, scaling, equating, diagnosis, and performance prediction were accomplished within this framework. The statistical software used to calibrate and scale the MI-ELPA was *Winsteps* Version 3.27 (Linacre & Wright, 2000).

7.1 The Rasch and Partial Credit Models

The most basic expression of the Rasch model is in the Item Characteristic Curve (ICC). It shows the probability of a correct response to an item as a function of the ability level. The probability of a correct response is bounded by 1 (certainty of a correct response) and 0 (certainty of an incorrect response). The ability scale is, in theory, unbounded. In practice, the ability scale ranges from -4 to +4 logits for heterogeneous ability groups.

The key step in the formulation and the point at which the Rasch dichotomous model merges with the Partial Credit Model (PCM), requires us to assume an additional response category. Suppose that, rather than scoring items as completely wrong or completely right, we add a category representing answers that, though not totally correct, are still clearly not totally incorrect. These relationships are shown in Figure 7.1.

The left-most curve ($j=0$) in Figure 7.1 represents the probability for all examinees getting a score of "0" (completely incorrect) on the item, given their ability. Those of very low ability (e.g., below -2) are very likely to be in this category and, in fact, are more likely to be in this category than the other two. Those receiving a "1" (partial credit) tend to fall in the middle range of abilities (the middle curve, $j=1$). The final, right-most curve ($j=2$) represents the probability for those receiving scores of "2" (completely correct). Very high-ability people are clearly more likely to be in this category than in any other, but there are still some of average and low ability who can get full credit for the item.

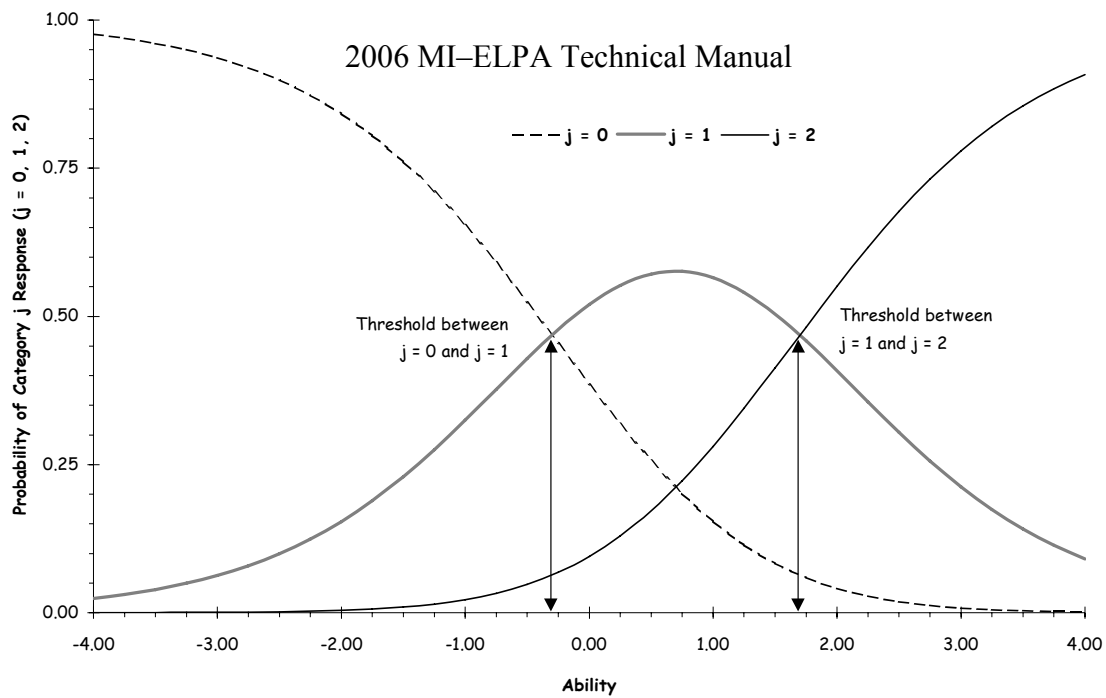


Figure 7.1: Category Response Curves for a Two-Step Item Using the PCM

An important implication of the formulation can be summarized as follows: If the commonly used Rasch model applied to dichotomously (right/wrong) scored items can be thought of as simply a special case of the PCM, then the act of scaling multiple-choice items together with polytomous items, whether they have three or more response categories, is a straightforward process of applying the measurement model. The quality of the scaling then can be assessed in terms of known procedures.

One important property of the PCM is its ability to separate the estimation of item/task parameters from the person parameters. With the PCM, as with the Rasch model, the total score given by the sum of the categories in which a person responds is a sufficient statistic for estimating a person's ability, i.e., no additional information needs be estimated. The total number of responses across examinees in a particular category is a sufficient statistic for estimating the step difficulty for that category. Thus, with PCM, the same total score will yield the same ability estimate for different examinees.

The PCM is a direct extension of the dichotomous one-parameter IRT model developed by Rasch in the 1950s (Rasch, 1980). For an item/task involving m_i score categories, one general expression for the probability of scoring x on item/task i is given by

$$P_{xi} = \exp \sum_{j=0}^x (\theta - D_{ij}) / \sum_{k=0}^{m_i} \left[\exp \sum_{j=0}^k (\theta - D_{ij}) \right] \quad (11)$$

Where $x = 0, 1, \dots, m_i$, and by definition,

$$\sum_{j=0}^0 (\theta - D_{ij}) = 0$$

The above equation gives the probability of scoring x on the i th test item as a function of ability (θ) and the difficulty of the m_i steps of the task (Masters, 1982).

According to this model, the probability of an examinee scoring in a particular category (step) is the sum of the logit (log-odds) differences between θ and D_{ij} of all the completed steps, divided by the sum of the differences of all the steps of a task. Thissen and Steinberg (1983) refer to this model as a divide-by-total model. The parameters estimated by this model are (a) an ability estimate for each person (or ability estimate at each raw score level) and (b) m_i threshold (difficulty) estimates for each task with $m_i + 1$ score categories. The item difficulty parameters are estimated using the Rasch model and the PCM discussed above and are provided in Appendix F of this report.

7.2 Calibration, Equating, and Scaling of the MI-ELPA

As part of the solution in using the Harcourt ELL item bank to meet the needs of the OEAA, Harcourt used the pre-existing SELP vertical scale together with some items from MWAC to create the MI-ELPA vertical scale. For the 2006 administration, the SELP items, which comprised the bulk of the items on the MI-ELPA, were fixed to the parameter values from the pre-existing vertical scale. That is, the SELP items were used as a common item link or anchor between the MI-ELPA and the SELP item bank. Any remaining non-SELP items on the MI-ELPA were calibrated together with the SELP items using the Rasch and Partial Credit models. Fixing the values of the SELP items prior to calibration resulted in the item difficulty and step parameters of all the items being placed on the same ability metric. The items were calibrated concurrently for all grade levels with the use of *Winsteps* 3.27 (Linacre, 2000). Although there was no “linkage” provided across grade levels, the vertical scale was maintained by the estimates of the SELP items that were used to place the new scale on the established SELP scale. The calibration estimates of item sets at each grade level were then used to obtain the raw score to theta scale.

The separate scales, one for each of the grade spans (i.e., K–2, 3–5, 6–8 and 9–12), and one for each of the strands within a grade span (i.e., Speaking, Listening, Reading, Writing, and Comprehension) were obtained by fixing (anchoring) the item parameters to their values estimated from the concurrent calibration. These item calibrations were then used to obtain the raw score-to-theta score tables for each of the four grade spans and the modalities within each grade span. Finally, when these calibrations and score tables were completed, the embedded field-test items for the 2006 administration were calibrated to the pre-existing vertical scale by using the core items as linking items.

A more detailed outline of the procedure follows:

- The calibration file was created from item-level data files using a sample that included Detroit, Dearborne, Grand Rapids, and the rest of the districts.
- The *Winsteps* 3.27 software program was used to conduct the item calibration by fixing the common SELP item parameters to their bank values.
- A comparison was made between the parameters from the initial calibration of the SELP items from the MI-ELPA and the parameters from the SELP item bank. Due to sampling error and scale indeterminacy, we did not expect the parameters for the two sets to be identical. However, we did expect the two sets of parameters to display a relatively clear linear relationship. (In fact, a linear relationship was found for the sets of item parameters at each of the four levels of the test).
- A second calibration was run, this time fixing the item parameters for the anchor set items to the SELP item bank values.
- The results of this second calibration were used as the operational item parameters used to create the final scales for the MI-ELPA spring 2006 administration.
- The final reporting scales were used to produce raw score-to-scale score conversion tables for the Total Test, and the Speaking, Listening, Reading, Writing, and Comprehension modalities (See section 7.5).

Appendices D.1–D.4 provide the raw-to-scale score conversion tables by grade span for the total test as well as by the Listening, Speaking, Reading, Writing, and Comprehension modalities. Braille form conversion tables are also provided for the two spans that were affected by changes made to the regular test, i.e., for grade spans 3–5 and 9–12. Similar tables for the Screener are provided in Appendices E.1–E.5, covering grade spans K, 1–2, 3–5, 6–8, and 9–12. The calibration data are representative of the population, covering the major districts in Michigan.

7.3 Vertical Scaling of SELP

An important component of any multilevel test is a continuous score scale that permits the interpretation of scores across levels of the test. According to Nitko (2004), a vertical scale is defined as an extended score scale that spans a series of levels and allows for the estimation of student growth along a continuum. In conducting the SELP multilevel equating, the adjacent levels of the test were scaled first, so that scores across levels were expressed on the same scale. The design that was utilized to obtain the vertical scale for the SELP was the common-person linking design, which is also referred to as the equivalent groups design (Kolen and Brennan, 2004, p. 389).

To accomplish the vertical scaling process, students in grades 3, 6, and 9 were involved. In the common-person linking design, the same students were administered two adjacent levels (on-level and one level lower) of the SELP. To control for test order and fatigue factors, a counterbalanced design was used to randomly administer the order of tests (lower level/higher level vs. higher level/lower level) to each participating classroom. Table 7.1 shows the research design for the vertical scaling of the SELP.

Table 7.1: Equating of Levels Research Design

Grade	SELP Off-Level		SELP On-Level	
	Level	Form	Level	Form
3	Primary	A	Elementary	A
6	Elementary	A	Middle	A
9	Middle	A	High	A

The *Winsteps* program was used to obtain Rasch item difficulties and person ability estimates for the two adjacent levels—Elementary/Primary, Middle/Elementary, and High/Middle. The adjacent levels were calibrated together; in other words, they were put on the same scale. Pairwised concurrent calibrations were conducted and level equating constants were calculated by applying the formula below:

$$K = \text{mean item difficulty Level(2)} - \text{mean item difficulty Level(1)} \quad (12)$$

A series of level equating constants were calculated and applied. The Elementary level constant was fixed at zero since it was chosen as the base scale, and then the level that was common between adjacent levels was used to calculate the level equating. The level equating constants for the Primary, Middle, and High were calculated using the formulas below:

$$K_{pe} = \mu_e - \mu_p$$

$$K_{me} = \mu_m - \mu_e$$

$$K_{he} = \mu_h - \mu_e,$$

where K_{pe} represents Constant (Primary/Elementary), K_{me} represents Constant (Middle/Elementary), K_{he} represents Constant (High/Elementary), and μ_e represents mean item difficulty (Elementary), etc.

Forms Equating

Maintaining continuity in the interpretation of results is essential for effectiveness of any large-scale assessment. One particularly effective technique for maintaining continuity of scores across years of administration of tests is to adopt a scale-score system for reporting results. Harcourt ensures that subsequent forms of the SELP are equated to the original Form A.

Test score information resulting from the Equating of Forms Program was used to develop scale scores for Form A and Form B. The scale scores indicate equivalent ability of students. To establish equivalence between forms, the *Winsteps* program was used to obtain Rasch item difficulties and person ability estimates. The two forms were treated as one extended test. This combined Rasch analysis placed both editions on the same common scale. Similarly, scale scores for Form A and Form C and Form A and Form D were developed.

A testing design similar to that of the Equating of Levels Program was utilized. Each student completed two forms of the SELP test. The order of administration of the two forms was counterbalanced by classroom to control for practice effects. To maintain the continuous vertical scale across forms, the scaling constants developed through the Equating of Levels Program were applied to test levels of each form.

Scale Scores

In addition to performance levels, SELP results are reported on a uniquely designed scale. Student raw scores, or the total number of points on the SELP, are converted into scale scores using a uniquely developed scaling procedure. The following equation was used to derive the scale scores:

$$SS = 35*(\theta) + 600 \quad (13)$$

In the above equation, θ was derived from item parameters that have been adjusted for test form and grade span/level.

The SELP scaling procedure involves linear transformations of the raw score points into scale score points. These transformations do not give more weight to particular subtests, and they change neither the rank ordering of students nor their performance-level classification. Linear transformation constants are utilized.

7.4 Linking MI-ELPA Scale to the SELP Vertical Scale³

As stated in Section 7.2, for the 2006 administration, the item parameters were fixed to the SELP item bank values. By fixing the known parameters of the common set of items, the items on the 2006 operational form were calibrated, the newly administered items were then located on the SELP scale. Once the scale locations of the 2006 MI-ELPA were known, IRT true score equating was used to relate the raw scores on the 2006 MI-ELPA to the SELP scale. In this process, the true score on the MI-ELPA with a given level of examinee ability is considered to be equivalent true score on the SELP associated with that level of examinee ability (Kolen and Brennan, 2004, p. 178).

³For additional details of how the original SELP vertical scale was established, please see the Stanford English Language Proficiency Test Technical Manual, 2005. Harcourt Assessment, Inc.

7.5 Scale Scores for the MI-ELPA

Once the MI-ELPA was linked to the SELP scale, the MI-ELPA raw scores were transferred to scale scores that ranged specifically between 300 to 801 on the total test, and 30 to 81 on each of the modalities, i.e., Speaking, Listening, Reading, Writing, and Comprehension.

The MI-ELPA scaling procedure involves linear transformations of the raw score points into scale score points. These transformations, like the SELP scale transformations described above, do not give more weight to particular subtests, and they change neither the rank ordering of students nor their performance level classification. Linear transformation constants are utilized. The equations used to establish each grade level and modality scores are summed in Table 7.2 below.

Table 7.2: Scale Score Transformation Equation for MI-ELPA Total Test and Modalities

Test/Modality	Scale Score Transfer Equation
Total Test	$31.25 * (\text{theta}) + 550$
Listening	$3.85 * (\text{theta}) + 57$
Speaking	$4.20 * (\text{theta}) + 57$
Reading	$3.75 * (\text{theta}) + 55$
Writing	$4.00 * (\text{theta}) + 52$
Comprehension	$3.55 * (\text{theta}) + 56$

7.6 Test Characteristic Curves for the MI-ELPA by Grade Span

Figures 7.2 below, display the test characteristic curves (TCCs) for the MI-ELPA by grade span. The TCCs are merely the average of the item response functions. As shown in the figure, the TCCs shift to the right as one progresses to the next higher level, indicating the relative increase in student ability required as one advances through the levels. This comparison is possible because of the vertical scale, whereby all test and student calibrations across grade spans are on the same scale.

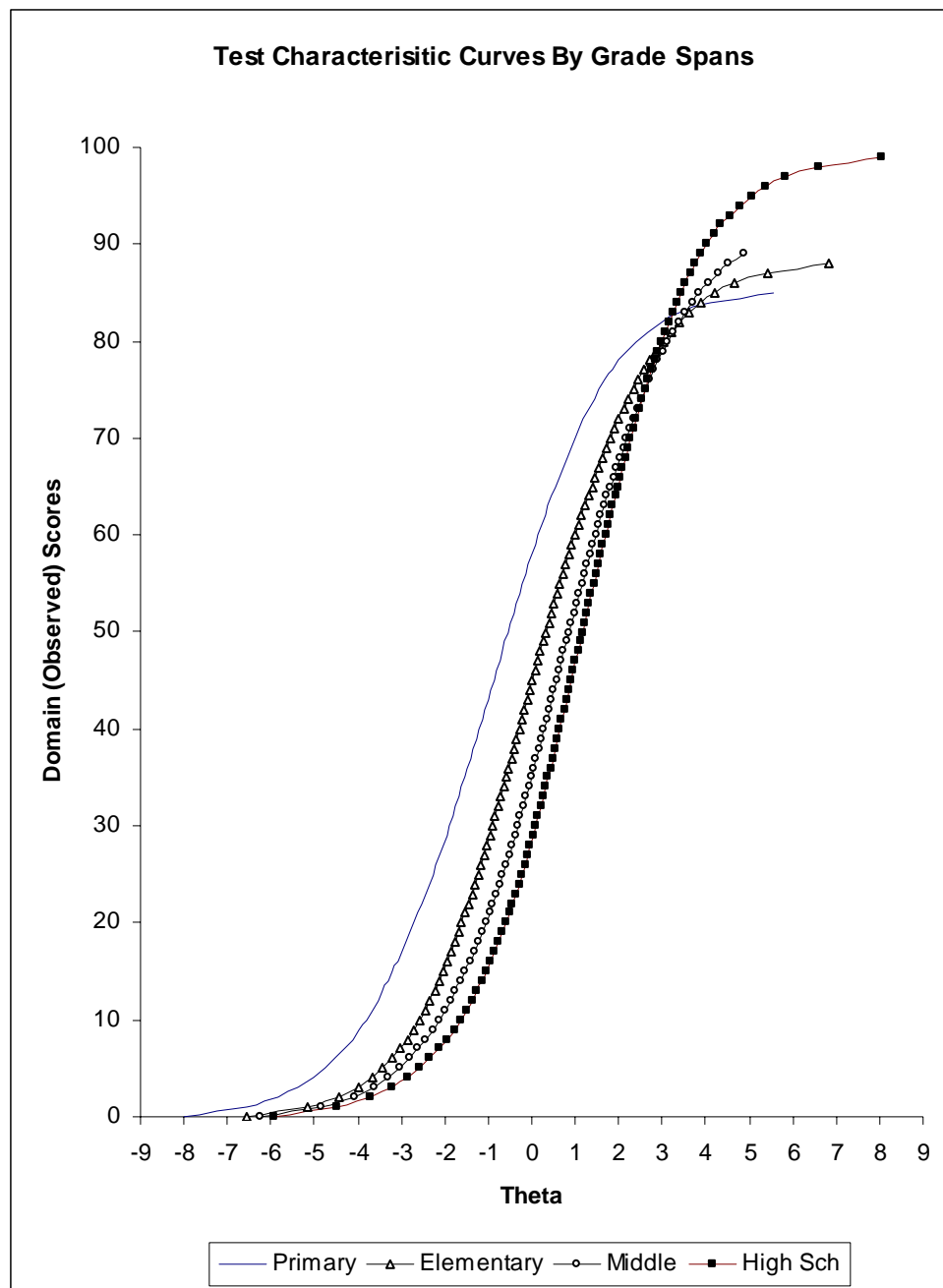


Figure 7.2: MI-ELPA Test Characteristic Curves (TCCs) by Grade Span

7.7 Linking Subsequent MI-ELPA Operational Tests across Years

Harcourt proposes to use IRT with internal common-item design for linking the MI-ELPA forms across years. The internal common items will be constructed using approximately 25 percent of the spring MI-ELPA.

Harcourt will use the pre-existing scale, a scale comparable to the SELP vertical scale, to create the MI-ELPA vertical scale. For example, for the 2007 administration, the linking items are the common items selected from the 2006 operational test. All non-linking items on the 2007 MI-ELPA will be calibrated together with the linking items using the Rasch and Partial Credit models. By fixing the values of the MI-ELPA items prior to calibration, this will result in the item difficulty and step parameters of all items being placed on the same ability metric.

SECTION 8. IRT STATISTICS

8.1 Model and Rationale for Use

In addition to reporting raw score summary statistics and item-level statistics using the classical test theory (CTT), the items on the MI-ELPA were also analyzed within the framework of Item Response Theory (IRT). The Rasch model (Rasch, 1960) for dichotomous items and the Partial Credit Model (Masters, 1982) for polytomous items were used for developing, scoring, and reporting the MI-ELPA. These models were recommended for several reasons.

First, the MI-ELPA vertical scale was created based on the pre-existing SELP vertical scale that was developed using the Rasch model. By using SELP items with known Rasch item difficulties, Harcourt was able to create the MI-ELPA vertical scale in a timely fashion.

Second, the sample size requirements for calibration, scaling, and equating under the Rasch and Partial Credit models are significantly smaller than for other IRT models. For example, the Rasch model requires on the order of 400 examinees per form for equating versus approximately 1,500 examinees per form under the 3PL IRT model (Kolen and Brennan, 2004, p. 288).

Finally, for the requirements of the MI-ELPA program, the Rasch model has one characteristic that makes it very useful. There exists a one-to-one relationship between raw scores and scale scores. That is, a student who answers a certain number of items correctly will receive the same scale score as a second student with the same raw score, regardless of which particular items within the test form were answered correctly. These reasons lead Harcourt to recommend that the Rasch model be adopted as the IRT methodology for the MI-ELPA.

8.2 Evidence of Model Fit

Fit statistics are used for evaluating the goodness-of-fit of a model to the data. Fit statistics are calculated by comparing the observed and expected trace lines obtained for an item after parameter estimates are obtained using a particular model. *Winsteps* provides two kinds of fit statistics called mean-squares that show the size of the randomness or amount of distortion of the measurement system.

The OUTFIT and the INFIT statistics are used in order to ascertain the suitability of the data for constructing variables and making measures with the Rasch model. These fit statistics are mean-square standardized residuals for item-by-person responses averaged over persons and partitioned between ability groups (OUTFIT) and within ability groups (INFIT). When the observed item characteristic curve (ICC) departs from the expected ICC from a reference value of 1, there is an expectation of high-ability students failing on an easy item or low-ability students succeeding on a difficult one. The OUTFIT mean-square evaluates the agreement between the observed ICC and the best fitting Rasch model curve over the ability sub-groups.

It is a standardized outlier-sensitive mean-square fit statistic, more sensitive to unexpected behavior by persons on items far from the person’s ability level. The INFIT, on the other hand, is a within-group mean-square, which summarizes the degree of misfit remaining within ability groups after the between-group misfit has been removed from the total. The INFIT, therefore, is a standardized information-weighted mean-square statistic, which is more sensitive to unexpected responses to items near the person’s ability level.

OUTFIT mean-squares are influenced by outliers and are usually easy to diagnose and remedy. INFIT mean-squares, on the other hand, are influenced by response patterns and are harder to diagnose and remedy. In general, mean-squares near 1.0 indicate little distortion of the measurement system, while values less than 1.0 indicate that observations are too predictable (redundancy, model overfit). Values greater than 1.0 indicate unpredictability (unmodeled noise, model underfit).

Englehard (1994) and other practitioners generally use 0.6 to 1.5 as the criteria for flagging deviations from the expected fit value of 1.00. Generally speaking, when item fit indices are lower than 0.6, they do not discriminate well and show a greater than expected degree of consistency. Similarly, a fit value higher than 1.5 indicates inconsistency in examinee scores on the item, i.e., some unexpectedly high scores are obtained by low-ability candidates, and low scores are obtained by high-ability candidates. Linacre and Wright, 1999, provide an overall guideline for evaluating mean-square fit statistics (see Table 8.1).

Table 8.1: Criteria to Evaluate Mean-Square Fit Statistics

Mean-Square	Interpretation
> 2.0	Distorts or degrades the measurement system
1.5 – 2.0	Unproductive for construction of measurement, but not degrading
0.5 – 1.5	Productive for measurement
< 0.5	Unproductive for measurement, but not degrading. May produce misleadingly good reliabilities and separations

Note. Adapted from Linacre & Wright, 1999.

In our analysis, items were only flagged if they distorted or degraded the measurement system, i.e., if they were > 2.0 logits. The OUTFIT and the INFIT statistics are presented by grade span in the item IRT statistics tables in Appendices F.1–F.4.

8.3 Rasch Statistics

Table 8.2 presents the grade span, the modality, the number of items in each modality, the maximum number of points attainable for each modality, and the average Rasch difficulty for each modality.

Table 8.2: Average Rasch Difficulty by Grade Span and Modality

Grade Span	Modality/Test	Number of Items	Max Points	Average Rasch Difficulty
K–2	Listening	21	21	-1.24
	Speaking	22	22	-1.04
	Reading	13	19	-0.46
	Writing	13	23	-2.11
	Comprehension	43	43	-0.98
	Total Test	69	85	-1.19
3–5	Listening	21	21	-0.12
	Speaking	21	21	1.12
	Reading	18	23	-0.12
	Writing	12	23	-0.49
	Comprehension	42	42	0.63
	Total Test	72	88	0.18
6–8	Listening	21	21	0.45
	Speaking	23	23	1.25
	Reading	17	23	0.55
	Writing	13	25	-0.34
	Comprehension	44	44	1.01
	Total Test	74	92	0.58
9–12	Listening	24	24	0.84
	Speaking	25	25	1.72
	Reading	18	25	1.30
	Writing	13	25	0.24
	Comprehension	49	49	1.24
	Total Test	80	99	1.12

Besides the INFIT and OUTFIT estimates, Appendices F.1–F.4 contain the results of the operational items for the MI-ELPA and include the Rasch item parameters. The following IRT item parameters are presented for each item, grouped by Listening/Speaking and Reading/Writing combinations:

- Number of students
- Rasch difficulty value
- Standard error of Rasch difficulty
- INFIT: Standardized information-weighted mean-square statistic, which is sensitive to unexpected behavior affecting responses to items near the person's ability level
- OUTFIT: Standardized outlier-sensitive mean-square fit statistic that is sensitive to unexpected behavior by persons on items far from the person's ability level

8.4 Item Information

Appendices H.1–H.13 provide item information at each of the three cut-scores. The information provided by item i about any point on the latent trait scale (θ) is defined mathematically as:

$$I_i(\theta) = \frac{[P'_i(\theta)]^2}{[P_i(\theta)][Q_i(\theta)]}, \quad (14)$$

where the numerator is the first derivative of $P_i(\theta)$. As specified by the equation, information is greater where the slope at a particular θ is greater, and the conditional variance at each ability level, θ . As Hambleton and Swaminathan (1996) state, “The greater the slope and smaller the variance, the greater the information, and hence the smaller the standard error of measurement.” (p. 105). For the Rasch model, the maximum information is constant and is obtained at a particular value on the ability scale. Items that provide the most information at the cuts would be considered for inclusion in form building.

SECTION 9. STANDARD SETTING

9.1 Introduction

The standard setting for the MI-ELPA was undertaken by Assessment and Evaluation Services in collaboration with Harcourt Assessment, Inc. The standard-setting sessions were conducted in Lansing, Michigan, from July 10 to July 12, 2006. The purpose of this meeting was to provide preliminary recommendations on performance cut-scores for the MI-ELPA.

For each of the four groups, there was one facilitator (a total of three from Harcourt and one from Assessment and Evaluation Services) to facilitate the technical part of the standard setting. In addition, a content specialist from Harcourt and an OEAA official together with a psychometrician from both Harcourt and OEAA were present to provide support during the standard-setting sessions. Data analysis was undertaken by a member of Assessment and Evaluation Services. Appendices G.1–G.7 provide detail information on the standard setting, including the agenda, the feedback provided by the panelists, and the targets for the modalities. This information was obtained from the files provided by Assessment and Evaluation Services.

9.2 Standard-Setting Methods

There are a variety of standard-setting methods, all of which require the judgments of educational experts and possibly other stakeholders. These experts are frequently referred to as judges, participants, or panelists (the term panelist will be used here). Acceptable methods for standard setting could be assessment-centered or student-centered (Jaeger, 1989). Assessment-centered methods focus panelists' attention on the items in the assessment. Panelists make decisions about how important and/or difficult the assessment content is and make judgments based on that importance. Student-centered methods focus panelists' attention on the actual performance of examinees or groups of examinees. Cut-scores are set based on student exemplars of different levels of competency. In addition, standards can be set using either a compensatory or conjunctive model (Hambleton & Plake, 1997). Compensatory models allow examinees who perform less well on some content to "make up for it" by performing better on other important content areas. Conjunctive models require that students perform at specified levels within each area of content.

Many standard-setting methods are better suited to specific conditions and certain item types. For example, the popular Modified Angoff method appears to work best with selected-response (SR) items (Cizek, 2001; Hambleton & Plake, 1997), while the "judgmental policy-capturing method" was designed specifically for complex performance assessments (Jaeger, 1995). Empirical research has repeatedly shown that different methods do not produce identical results, and it is important to consider that many measurement experts no longer believe that "true" cut-scores exist (Zieky, 2001).

Therefore, it is crucial that the method chosen meets the needs of the testing program. *The Standards for Educational and Psychological Testing* (AERA, APA, NCME, 1999) details issues that should be addressed in all educational testing situations. While not specifically addressing standard setting, several standards are relevant.

- Standard 4.19—“When proposed score interpretations involve one or more cut scores, the rationale and procedures used for establishing should be clearly documented.”

Standard 4.19 states the purpose of this manual and recommends its content. This manual will document the reason for standard-setting methods and clearly describe them. This will include the methods used and rationale for those procedures. This manual will also provide the results of the standard setting and an estimate of variation of cut-scores relevant to the replication of the process.

- Standard 4.20—“When feasible, cut scores defining categories with distinct substantive interpretations should be established on the basis of sound empirical data concerning the relation of test performance to relevant criteria.”

Although Standard 4.20 may be focused on employment testing where distinct categories have been established and the basis for the criterion can be empirically demonstrated, the discussion of the standard does state that “a carefully designed and implemented procedure based solely on judgments of content relevance and item difficulty may be preferable to an empirical study.” In the case of a content-based assessment, the judgments of panelists according to performance-level descriptors take the place of empirical data.

- Standard 4.21—“When cut scores defining pass-fail or proficiency categories are based on direct judgments about the adequacy of items or test performances or performance levels, the judgmental process should be designed so that judges can bring their knowledge and experience to bear in a reasonable way.”

Standard 4.21 states the need for standard-setting methods to provide panelists with reasonable judgment tasks based on their experiences. In both the Item Mapping and Body of Work methods, panelists are asked to think about student performance in reference to the performance-level descriptors. This task is done by teachers every day in the classroom. These methods are the result of a refinement of standard-setting methods so that they can better meet the requirements of Standard 4.21.

9.3 Standard-Setting Model and Process

Item mapping is a well-established method available for establishing performance standards. The item-mapping procedure is capable of incorporating both multiple-choice and constructed-response items into the same process (Mitzel, H.C., Lewis, D.M., & Green, D.R., 2001). It has several other favorable characteristics, including:

- Simplifying the judgment task by reducing the cognitive load required by panelists
- Connecting the judgment task of setting cut-scores with the measurement model
- Connecting content with performance-level descriptors

The Item Mapping procedure/bookmarking used for setting the MI-ELPA cut-scores required the panelists to make judgments about student performance defined by the Performance Level Descriptors (PLDs or Standards). (See the display of PLDs in Appendix G.4.) The task is a series of judgments about how students just at the standard will perform on the test items. To make the task more easily accomplished, the test items had been arranged in a booklet by their difficulty. The easiest item was on the first page, and the most difficult item was on the last page. Essentially this process allows multiple-choice and open-ended items to be judged in the same manner. Assessment items are arranged or mapped in order of difficulty, and panelists make decisions about performance of students according to the definitions. See Appendix G.5 for an example of how items are placed in ascending order of difficulty before being placed in the booklet. Panelists had to decide along a continuum of item difficulty how a particular set of students just meeting the definition will perform. Essentially, panelists were selecting along the continuum of items where a certain percentage (0.50 and 0.67 are most commonly used) would answer an item correctly but the same percentage would not answer the next-hardest item correctly. As shown in Appendix G.5, for the MI-ELPA standard setting, it was decided that the bookmark location for arranging items according to their difficulty would be 0.67 probability of obtaining the item score. IRT scaling methods allowed the scaling of the assessment items and open-ended item levels so that panelists' decisions could be translated into an ability level and a raw-score equivalent on the assessment.

Panelists set cut-scores based on 100 hypothetical “borderline” students; therefore they had to think about the characteristics that defined this population. In working on the PLDs, they had outlined what students at each level should know and be able to do, and in item mapping panelists took that information and adapted it to developing cut-scores to distinguish students across the four levels. For example, as shown in the PLDs depicted in Appendix G.4, for a student in Grade 1, the Proficient level in the Listening modality indicates that the student must follow simple and complex directions and listen and respond to stories, texts, and social interactions appropriately. The differentiating factor between the Proficient level and the Intermediate B level is that the same descriptors for the Proficient level must also be followed by students for placement in the Intermediate B level with one exception, i.e., for the Intermediate level the descriptors must be followed *most of the time* as opposed to the regular expectations of the Proficient level. Once this information was obtained, panelists performed item mapping, setting a cut for each performance level between an item that 67 % of the students would answer correctly and the next most difficult item, which 67% of the students would not answer correctly.

The standards that were recommended will become part of a larger set of standards used by the state to describe the results of the assessment system. These recommendations need to be made as a system of standards that educators and the public will use to evaluate student, school, district, and state performance.

9.4 Committees of Panelists

Four standard-setting committees were established to set the cut-scores for the four grade spans of the MI-ELPA. As indicated in Table 9.1, the first group recommended standards on grades K–2, the second group recommended standards on grades 3–5, the third group recommended standards on grades 6–8, and the fourth group recommended standards on grades 9–12.

The panelists were all ESL teachers or specialists. Approximately six panelists had experience in the grade range, and two panelists spanned the other ranges, one above and one below where possible.

An attempt was made to obtain panelists who work with different languages. They were sampled from the state based on the frequency of students in ESL programs. Geographic diversity was based on ESL program areas and not the entire state.

Table 9.1: Panel Composition for Standard-Setting Committees

Grade	Group	Number of Panelists
K 1 2	1	8
3 4 5	2	8
6 7 8	3	8
9 10 11 12	4	8

9.5 Performance Levels and Cut-Scores

For the MI-ELPA, four performance levels, which correspond to three cut scores, are required. The four performance levels are:

- Beginning
- Intermediate A
- Intermediate B
- Proficient

The three cut scores are:

- Intermediate A (between the Beginning and Intermediate A performance levels)
- Intermediate B (between the Intermediate A and Intermediate B performance levels)
- Proficient (between the Intermediate B and Proficient performance levels)

To set the three cut points, the item-mapping procedure was utilized. The standard-setting process is briefly described below.

9.6 The Use of the Vertical Scale

The ELPA is a new assessment and no previous standards exist. However, the ELPA has a vertical scale that allows comparison across the 4 levels of the assessment. The ELPA scale was developed using embedded items from the Stanford English Language Proficiency (SELP) assessment developed and published by Harcourt.

An important component of any multilevel test is a continuous score scale that permits the interpretation of scores across levels of the test. This is carried out for adjacent levels, so that scores across levels are expressed on the same scale.

Harcourt’s research study to establish the vertical scaling for SELP involved students in grades 3, 6, and 9. Students were administered two adjacent levels (on-level and one level lower) of SELP. To control for test order and fatigue factors, a counterbalanced design was used to randomly administer the order of tests (lower level/higher level vs. higher level/lower level) to each participating classroom. Through the scaling study, equating constants for each level as well as the scaling intercept and slope were derived. To link a customized assessment to the SELP vertical scale, the user would apply the level equating constants first if the items are selected from different SELP levels. Then, as soon as the theta values (Rasch difficulties) are available after calibration, the vertical scale intercept and slope would be applied to the theta values to generate the raw- to-scale-score table.

Test score information resulting from the Equating of Forms Program was used to develop scale scores for Form A and Form B. The scale scores indicate an equivalent ability of students. To establish equivalence between forms, the *Winsteps* program was used to obtain Rasch item difficulties and person ability estimates. The two forms were treated as one extended test. This combined Rasch analysis placed both editions on the same common logistic scale. The data were also used to establish the alternate forms reliability of the tests. A testing design similar to that of the vertical scaling was utilized.

Each student completed two forms of SELP. The order of administration of the two forms was counterbalanced by classroom to obviate practice effects. To maintain the continuous vertical scale across forms, the scaling constants developed through the Equating of Levels Program were applied to test levels of each form.

In addition to performance levels, SELP results are reported on a uniquely designed scale. Student raw scores, or the total number of points on the SELP, are converted into scale scores using a uniquely developed scaling procedure. The SELP scaling procedure involves linear transformations of the raw score points into scale score points. These transformations do not give more weight to particular subtests, and they change neither the rank ordering of students nor their performance-level classification. Linear transformation constants are utilized in the process. A vertical scale for the MI-ELPA will be developed based on the inclusion of test items from SELP into the MI-ELPA. These items are also used to measure language skills and are part of the total score on the MI-ELPA.

The standard-setting committees used the vertical scale so that a logical system of standards could be set across the grades and levels. The four committees set standards at grades 2, 3, 8, and 9 as the first set of grades, then moved on to the second set of grades (grades K, 5, 6, and 12), and finally to grades 1, 4, 7, 9 and 10. The vertical scale was used to inform the committees as they set the standards. The vertical scale information was provided at the end of Round 2 with the impact data.

9.7 Standard-Setting Process

The standard setting began with introductions from the OEAA, Harcourt, and panelists. This was followed by a presentation by the lead facilitator on the role of the panelists in the standard-setting process, setting performance standards, and placing cut scores. The goal was to familiarize panelists with the standard-setting process and the item-mapping procedure. This session took place in a large group setting (all four groups together).

After the orientation, the panelists were separated into specific breakout room according to their group assignments. Each group/room was led by a facilitator who is an expert in the standard-setting methodology, and assessment specialists rotated from group to group in order to provide content support. In addition, the panel members were further divided into three smaller table groups within their grade spans, each composed of five to seven members. These small groups worked independently but had the opportunity to collaborate with the other table groups in their grade span during the standard-setting process. The following sequences of tasks were completed.

Review of the Assessment

Their first task was to review the assessment blueprint. This was done in order for the panelists to gain an understanding of what the assessment is intended to measure. Discussions about the assessment content, the use of different item types, and number of questions were conducted. The panel members further defined the general performance-level descriptors into specific descriptors to help the panel members come to a shared understanding about what it meant to be performing at each of the performance levels. The facilitator led this discussion with support from the assessment specialists who floated between the rooms.

Experiencing the Assessment

Next, the panel members had an opportunity to experience the assessment administered at the grade span assigned to them. This was an effective way to demonstrate to the panelists the knowledge and skills that students must possess to obtain a high score. It is assumed that panelists are likely to set more realistic performance standards if they experience the assessment themselves.

Scoring the Assessment

After the panelists finished taking the assessment, they were provided with an answer key to grade their tests. The panelists scored their own assessments using the scoring rubrics and answer key provided. The scoring process offered an opportunity for the panelists to develop an understanding of the scoring of open-ended responses. They were provided with exemplars of score points. A discussion session then followed the scoring of the assessment.

Review of Student Performance Levels

Panelists reviewed the previously established definitions of performance levels (Appendix G.3). Then they discussed the performance levels. The goal was to help panelists clearly distinguish between student performance levels. Panelists' suggestions were related to the performance standards and content frameworks. The suggestions were retained for reference during the standard-setting process. Panelists reviewed definitions and offered illustrative suggestions for the Beginning, Intermediate A, Intermediate B, and Proficient performance standards. After all the performance levels were reviewed, a discussion session was held. The focus was on the characteristics and interrelationships between and among performance standards.

Three Rounds of Ratings

The actual standard setting proceeded in three rounds. Each round was designed to foster increased consensus among panelists, although reaching consensus was not necessary. Panelists expressed their cut-score judgment by placing a marker on the item that a student at that threshold of a performance level should master. One marker was placed for each cut score. There were three cut scores.

During the Round 1 ratings, each panelist began by setting his/her three cut scores. The data were captured for each panelist. Before the Round 2 ratings, panelists were provided feedback on the Round 1 cut-score positions of all panelists and the median cut-scores of their group. The panelists then discussed the Round 1 results. After the discussions, the Round 2 cuts were made, followed by further discussions. At this point, the panelists were provided with information about the percentage of students who would be classified in each of the performance levels, if those cuts were to be implemented. These percentages were based on all students who took the assessment in spring 2006. An example of the format of the information provided to panelists at the end of Round 2 is depicted in Appendix G.2.

In order to promote consistency across the grade spans, the groups came together to discuss the process and results of their assigned grades between all grade spans. Panelists then returned to their breakout groups and proceeded to make their Round 3 ratings. The median cut-scores of the panelists then served as the starting point for the decision-makers on establishing the cut-scores for the assessment.

Evaluation

At the end of the final rating, panelists filled out an evaluation form that assessed their beliefs about each component of the standard-setting process and how confident they felt in the overall results (Appendix G.5). After the evaluation the panelists had a debriefing session.

9.8 Agendas

Panelists completed three rounds of standard setting for each grade over a three-day period. Grade-level standards were completed in three sets:

- Set 1 established standards for grades 2, 3, 8, and 9 and followed the first-day agenda in Appendix G.1.
- Set 2 set standards for four grades (grades K, 5, 6, and 8) and followed the second-day agenda in Appendix G.1.
- Set 3 covered the remaining grades: grades 1, 4, 7, and 10 and 11. Because the High School group had four grade levels, the grades 10 and 11 standards were done together during the third set and followed the third-day agenda.

For a complete listing of the agenda, refer to Appendix G.1.

9.9 Summary Statistics for the Three Rounds of Ratings

Panelists completed three rounds of standard setting for each grade over a three-day period. Grade-level standards were completed in three sets. Set 1 was grades 2, 3, 8 and 9. This allowed the Primary–Grade 2 committee to meet with the Elementary–Grade 5 committee after Round 2. The Middle–Grade 8 committee met with the High School committee after Round 2 of Set 1.

Set 2 was grades K, 5, 6, and 8. The Elementary–Grade 5 committee met with the Middle–Grade 8 committee after Round 2 of Set 2. Set 3 covered the remaining grades: grades 1, 4, 7, and 10 and 11. Because the High School group had four grade levels, the grades 10 and 11 standards were done together during the third set. The following tables (Table 9.2–9.5) show the raw score standard for each round by grade.

Table 9.2: Primary School Level Raw Score Standards by Rounds

Grade	Round	Proficiency-Level Cuts		
		INTA	INTB	PROF
K	1	32	39	46
	2	31	37	44
	3	31	42	49
1	1	42	55	66
	2	43	54	68
	3	43	54	68
2	1	44	62	75
	2	47	61	74
	3	47	60	74

Note. INTA = Intermediate A. INTB = Intermediate B. PROF = Proficient.

Table 9.3: Elementary School Level Raw Score Standards by Rounds

Grade	Round	Proficiency-Level Cuts		
		INTA	INTB	PROF
3	1	31	50	66
	2	31	51	69
	3	32	52	71
4	1	34	52	73
	2	34	53	73
	3	34	55	73
5	1	44	64	79
	2	37	58	76
	3	38	58	75

Table 9.4: Middle School Level Raw Score Standards by Rounds

Grade	Round	Proficiency-Level Cuts		
		INTA	INTB	PROF
6	1	36	59	71
	2	37	62	75
	3	37	61	76
7	1	39	65	78
	2	39	65	78
	3	39	65	78
8	1	46	65	80
	2	42	64	80
	3	43	66	80

Table 9.5: High School Level Raw Score Standards by Rounds

Grade	Round	Proficiency-Level Cuts		
		INTA	INTB	PROF
9	1	50	71	82
	2	46	68	82
	3	49	69	85
10	1	50	70	85
	2	51	70	86
	3	51	70	86
11	1	52	75	87
	2	52	75	87
	3	52	75	87
12	1	62	78	94
	2	56	78	93
	3	54	78	89

The *Standard Setting Results* section in Appendix G.6 provides the summary statistics for the round-by-round results by grade by the three performance-level cuts. The tables show the raw score standard for each round by grade.

The standard setting resulted in the recommendation of three cut-scores (Intermediate A, Intermediate B, and Proficient) across 13 grade levels. The graph below indicates the percent of students from the spring sample that would fall into each of the four categories (Basic, Intermediate A, Intermediate B, and Proficient) given the standards that were recommended by the committees. Caution should be used in interpreting the percent of students in each category. Unlike the census MEAP, this is a sample of students and not the entire population. There may be factors that bias selection of students for testing by grade or language ability. Although the Percent in Category information was examined by panelists during the standard setting, it was not the primary focus of discussion. The level of standards in reference to the vertical scale was a more important tool for evaluating the standards, and the panelists focused more on those numbers.

9.10 Evaluation Results

Panelists completed an evaluation form at the conclusion of the standard-setting meeting. They were asked about the process, the steps in the process, the facilities, and their confidence in the standards they had set. Appendix G.5 provides the results of the panelists' feedback. A tally of each committee's responses is presented. The forms indicate how many panelists responded to each category.

In general the feedback reflects satisfaction with the process and confidence in the standards that were recommended.

9.11 Post-Standard-Setting Analyses

The median scores from the standard-setting committees were used as the recommended cuts. The cut-scores were based on the total MI-ELPA score. After the standard-setting meetings, several post-standard-setting analyses were performed. The first step was to look up the equivalent scale scores corresponding to the raw-score cuts recommended by the committees. Graphs were then plotted using the grades as the independent variable and scale score as the dependent variable. The three cut-scores were then plotted on the same graph to show that the cuts were monotonically increasing from the lower cuts to the higher cuts. As stated earlier, the percentage of students falling into each of the performance levels was calculated for each grade should those cut points be adopted. Impact information, i.e., the percentage of students falling into each of the performance levels, was provided to the OEAA to make their final decisions on the cut-scores for the MI-ELPA.

9.12 Final Performance-Level Cut-Scores for the MI-ELPA

Table 9.6 contains the vertical scale values for the standards recommended at the end of Round 3. Vertical scale values are increasing for each grade. This is consistent with the concept that as students move up the grades the English language ability that describes the categories should increase.

Table 9.6: Final Performance-Level Cut-Scores

Grade	Total MI-ELPA								
	Raw Score			Scale Score			Theta		
	INTA	INTB	PROF	INTA	INTB	PROF	INTA	INTB	PROF
K	31	42	49	493	517	531	-1.82	-1.07	-0.61
1	43	54	68	519	541	575	-1.01	-0.28	0.80
2	47	60	74	527	555	595	-0.74	0.15	1.43
3	32	52	71	531	572	619	-0.60	0.71	2.22
4	34	55	73	535	579	626	-0.46	0.91	2.42
5	38	58	75	544	585	633	-0.21	1.13	2.65
6	37	61	76	554	598	635	0.11	1.54	2.71
7	39	65	78	557	607	641	0.23	1.81	2.91
8	43	66	80	564	609	648	0.46	1.88	3.14
9	49	69	85	585	619	658	1.12	2.22	3.44
10	51	70	86	588	621	661	1.22	2.28	3.54
11	52	75	87	590	632	664	1.27	2.61	3.66
12	54	78	89	593	638	672	1.38	2.83	3.90

Note. INTA = Intermediate A; INTB = Intermediate B; PROF = Proficient

The final cut-scores adopted by OEAA for the 2006 administration of the MI-ELPA for the test in raw score points, scale score, and theta metric were the same as those recommended by the standard-setting committee. There are three cut-scores that correspond to four performance levels. Any score below the Intermediate A cut-score falls into the Beginning performance level.

9.13 Calculation of Achievement “Targets” for Each Modality

Achievement “targets” for each of the five modalities, i.e., Listening, Speaking, Reading, Writing, and Comprehension, were set by calculating the average raw score for each of these modalities of those students who received a score equal to or greater than the proficiency cut for the total test. These targets are provided in Appendix G.8.

9.14 Calculation of the Performance-Level Cuts for the Screener

The same performance-level theta cut set for the total test was used to set the performance-level cut for the screener. The total number of items for the screener together with their theta and raw cuts are provided in Appendix G.9.

SECTION 10. SUMMARY OF OPERATIONAL TEST RESULTS

This section presents both the raw score and scale score summaries for each of the modalities and for the total MI-ELPA by grade. Table 10.1 presents the raw-score summary by grade. Table 10.2 presents the scale-score summary by grade. Tables 10.1 and 10.2 include the sample size, mean, median, interquartile range, and standard deviation. Table 10.3 presents the percentage of students in each of the proficiency levels by grade.

Table 10.1: Raw-Score Summary by Grade, Modality, and Total Test

Grade	Test	N-Count	Mean	Median	IQR	SD
K	Listening	7773	12.00	12	4	2.80
	Speaking	7773	15.34	16	7	4.81
	Reading	7773	9.48	9	5	3.31
	Writing	7773	4.60	4	5	3.74
	Comprehension	7773	16.31	16	5	3.90
	Total Test	7773	41.42	41	14	10.91
1	Listening	7507	14.29	14	3	2.72
	Speaking	7507	18.44	19	4	3.94
	Reading	7507	14.32	14	5	3.49
	Writing	7507	11.97	13	5	4.24
	Comprehension	7507	21.22	21	5	4.18
	Total Test	7507	59.02	61	14	11.45
2	Listening	6805	16.19	17	3	2.62
	Speaking	6805	19.94	21	3	3.42
	Reading	6805	17.42	18	5	3.47
	Writing	6805	15.02	16	3	3.52
	Comprehension	6805	24.98	25	6	4.34
	Total Test	6805	68.58	71	12	10.70
3	Listening	6116	16.17	17	4	3.31
	Speaking	6116	19.47	21	4	3.80
	Reading	6116	11.97	12	6	3.94
	Writing	6116	16.94	18	5	3.90
	Comprehension	6116	23.66	24	8	5.67
	Total Test	6116	64.55	67	14	12.32
4	Listening	5468	17.03	18	3	3.14
	Speaking	5468	20.05	21	3	3.60
	Reading	5468	13.65	14	6	3.93
	Writing	5468	18.14	19	4	3.63
	Comprehension	5468	25.90	27	7	5.66
	Total Test	5468	68.87	72	13	11.93
5	Listening	5200	17.69	18	3	2.92
	Speaking	5200	20.38	21	3	3.45
	Reading	5200	14.89	16	5	3.87
	Writing	5200	18.92	20	3	3.35
	Comprehension	5200	27.59	29	7	5.46
	Total Test	5200	71.88	75	11	11.45

Table 10.1: Raw-Score Summary by Grade, Modality, and Total Test (Continued)

Grade	Test	N-Count	Mean	Median	IQR	SD
6	Listening	4646	15.68	16	4	3.15
	Speaking	4646	21.58	23	3	4.32
	Reading	4646	14.49	15	7	4.27
	Writing	4646	16.56	17	4	3.90
	Comprehension	4646	21.33	22	7	5.06
	Total Test	4646	68.31	71	15	13.12
7	Listening	4164	16.01	17	3	3.13
	Speaking	4164	21.59	23	3	4.44
	Reading	4164	15.10	16	6	4.30
	Writing	4164	16.99	18	5	3.87
	Comprehension	4164	22.01	23	7	5.06
	Total Test	4164	69.69	73	14	13.30
8	Listening	3830	16.32	17	4	3.18
	Speaking	3830	21.85	23	4	4.34
	Reading	3830	15.85	17	6	4.28
	Writing	3830	17.45	18	4	3.83
	Comprehension	3830	22.81	24	7	5.15
	Total Test	3830	71.47	75	13	13.41
9	Listening	3967	17.33	18	6	4.42
	Speaking	3967	20.69	22	5	5.11
	Reading	3967	16.43	18	9	5.33
	Writing	3967	16.19	17	7	4.86
	Comprehension	3967	22.60	24	8	5.82
	Total Test	3967	70.65	75	23	17.20
10	Listening	2899	18.04	19	5	4.20
	Speaking	2899	21.38	23	4	4.18
	Reading	2899	17.54	19	8	5.05
	Writing	2899	17.20	18	5	4.32
	Comprehension	2899	23.70	25	7	5.54
	Total Test	2899	74.16	78	19	15.35
11	Listening	1984	18.44	19	5	3.92
	Speaking	1984	21.61	23	4	3.90
	Reading	1984	18.03	19	7	4.78
	Writing	1984	17.68	18	5	4.05
	Comprehension	1984	24.26	25	7	5.20
	Total Test	1984	75.75	79	18	14.14
12	Listening	1494	18.89	20	5	3.84
	Speaking	1494	21.95	23	4	3.57
	Reading	1494	18.57	20	6	4.80
	Writing	1494	18.06	19	5	4.08
	Comprehension	1494	24.87	26	7	5.21
	Total Test	1494	77.47	80	17	13.77

Note. 1. The total n-count for each grade was obtained after deleting all raw scores of 999 while “Omits” and “Blanks” were scored as “0s.” 2. IQR = Interquartile Range.

Table 10.2: Scale-Score Summary

Grade	Test	N-Count	Mean	Median	IQR	SD
K	MI-ELPA	7773	514.94	514	29	24.24
1	MI-ELPA	7507	554.89	557	33	28.18
2	MI-ELPA	6805	605.18	608	38	32.26
3	MI-ELPA	6116	618.38	622	41	33.88
4	MI-ELPA	5468	640.31	640	40	40.22
5	MI-ELPA	5200	628.95	633	40	35.18
6	MI-ELPA	4646	618.50	621	39	31.38
7	MI-ELPA	4164	622.40	626	38	32.64
8	MI-ELPA	3830	628.02	632	37	34.14
9	MI-ELPA	3967	627.99	632	49	35.96
10	MI-ELPA	2899	635.85	638	44	34.48
11	MI-ELPA	1984	639.54	641	43	33.65
12	MI-ELPA	1494	644.03	643	45	34.12

Note. 1. The total n-count for each grade was obtained after deleting all raw scores of 999 while “Omits” and “Blanks” were scored as “0s.”

2. Generally speaking, the mean for each grade should increase from one grade to the next higher grade in a similar manner as shown in Table 9.6 of this manual, which depicts increases across grade levels. However, due to artifacts of the population whereby some grades may have a greater percentage of Higher scoring students than the next higher grade, the mean for the lower grade can be higher than the next higher grade/s.

Table 10.3: Percent of Students in Each Proficiency Level by Grade

Grade	Test	N- Count	Proficiency Levels			
			1	2	3	4
K	MI-ELPA	7773	14.92	23.92	36.28	24.88
1	MI-ELPA	7507	8.75	49.38	18.13	23.74
2	MI-ELPA	6805	4.98	49.01	9.39	36.62
3	MI-ELPA	6116	2.65	51.78	10.12	35.45
4	MI-ELPA	5468	2.41	43.40	7.68	46.51
5	MI-ELPA	5200	2.31	39.12	7.06	51.52
6	MI-ELPA	4646	3.59	46.84	16.81	32.76
7	MI-ELPA	4164	4.11	45.51	20.75	29.63
8	MI-ELPA	3830	4.99	47.75	17.86	29.40
9	MI-ELPA	3967	13.13	43.13	21.88	21.86
10	MI-ELPA	2899	9.38	45.05	20.80	24.77
11	MI-ELPA	1984	7.01	37.40	31.80	23.79
12	MI-ELPA	1494	6.89	36.14	34.14	22.82

Note. The total n-count for each grade was obtained after deleting all raw scores of 999 while “Omits” and “Blanks” were scored as “0s.”

Table 10.4: Percent of Students in Each Proficiency Level by Grade and Modality

Grade	Modality	N-Count	Achievement Target - % of Proficient Students
K	Listening	7773	29.77
	Speaking	7773	20.08
	Reading	7773	36.40
	Writing	7773	26.82
	Comprehension	7773	28.07
1	Listening	7507	21.11
	Speaking	7507	21.62
	Reading	7507	26.19
	Writing	7507	20.97
	Comprehension	7507	21.06
2	Listening	6805	34.40
	Speaking	6805	38.71
	Reading	6805	32.06
	Writing	6805	20.87
	Comprehension	6805	40.26
3	Listening	6116	39.94
	Speaking	6116	33.14
	Reading	6116	38.80
	Writing	6116	26.95
	Comprehension	6116	34.52
4	Listening	5468	54.02
	Speaking	5468	42.19
	Reading	5468	36.85
	Writing	5468	42.54
	Comprehension	5468	37.51
5	Listening	5200	48.27
	Speaking	5200	47.60
	Reading	5200	51.50
	Writing	5200	36.87
	Comprehension	5200	52.29
6	Listening	4646	31.83
	Speaking	4646	39.07
	Reading	4646	36.29
	Writing	4646	35.51
	Comprehension	4646	29.57
7	Listening	4164	36.36
	Speaking	4164	40.47
	Reading	4164	24.71
	Writing	4164	26.80
	Comprehension	4164	26.92
8	Listening	3830	25.59
	Speaking	3830	26.63
	Reading	3830	22.66
	Writing	3830	33.24
	Comprehension	3830	26.16
9	Listening	3967	16.54
	Speaking	3967	35.27
	Reading	3967	19.08
	Writing	3967	28.28

2006 MI-ELPA Technical Manual

Grade	Modality	N-Count	Achievement Target -
			% of Proficient Students
10	Comprehension	3967	29.42
	Listening	2899	22.97
	Speaking	2899	39.46
	Reading	2899	16.70
	Writing	2899	34.56
11	Comprehension	2899	28.22
	Listening	1984	23.99
	Speaking	1984	39.97
	Reading	1984	18.75
	Writing	1984	26.46
12	Comprehension	1984	30.09
	Listening	1494	16.67
	Speaking	1494	43.04
	Reading	1494	23.76
	Writing	1494	19.34
	Comprehension	1494	28.58

Note. The total n-count for each grade was obtained after deleting all raw scores of 999 while “Omits” and “Blanks” were scored as “0s.”